Among Digitized Manuscripts

Handbook of Oriental Studies

Handbuch der Orientalistik

SECTION ONE

The Near and Middle East

Edited by

Maribel Fierro (*Madrid*)
M. Şukru Hanioğlu (*Princeton*)
Renata Holod (*University of Pennsylvania*)
Florian Schwarz (*Vienna*)

VOLUME 137

The titles published in this series are listed at *brill.com/ho1*

Among Digitized Manuscripts

Philology, Codicology, Paleography in a Digital World

By

L.W.C. van Lit, O.P.

BRILL

LEIDEN | BOSTON

The digital resources referred to in this book are freely available online at http://doi.org/10.5281/zenodo.3371200.

Cover illustration: Manuscript page from the *ʿAmal al-munāsakhāt bi-al-jadwal* by Aḥmad ibn Muḥammad Ibn al-Hāʾim.

The Library of Congress Cataloging-in-Publication Data is available online at http://catalog.loc.gov
LC record available at https://lccn.loc.gov/2019037402

Typeface for the Latin, Greek, and Cyrillic scripts: "Brill". See and download: brill.com/brill-typeface.

ISSN 0169-9423
ISBN 978-90-04-41521-8 (hardback)
ISBN 978-90-04-40035-1 (e-book)

This book is printed on acid-free paper and produced in a sustainable manner.

Dedicated to my brethren
Theo, Jan, Richard, Stefan, Michael-Dominique, and Augustinus

Wise men and fools must both perish,
and leave their wealth to others

Contents

Acknowledgments

This book is the result of years of slowly piecing together experience. I should note here that I have experience with coding and designing from a young age. In the final year of elementary school in 1998, I created the school's website, and throughout my teens I made Flash animations. What also helped was that I studied mathematics in college. Coming with that background into Islamic studies obviously set me up to engage with digital humanities. In 2013, I started a weblog called *The Digital Orientalist*, with the intent to share my workflows and homemade hacks and tools that make life a little easier for someone in Islamic studies using a computer. Originally, my idea was to write posts that could ultimately function as paragraphs in a book-length introduction to the role of computers and digital resources for students in the humanities. As I saw it, and still is the case, students may learn research-related methodologies, and may learn how to read a text or write an essay, but rarely is there formal training in how the computer is integrated into all of this.

In 2015, I participated in various workshops and events at the DH Lab of Yale University, which helped me refine and reorient my endeavors. In 2016, I focused on working exclusively with digitized manuscripts. In the spring of 2017 I had prepared a more formal investigation of what it means to work with digitized manuscripts versus actual manuscripts which I presented at Freie Universität Berlin. My research stay was sponsored by the Dahlem Humanities Center, on invitation of Olly Akkerman. In the summer of that year, I had expanded my work into two separate articles on which I presented at Jyväskylä University, in Finland. I worked there as a postdoc in the ERC-funded project 'Epistemic Transitions in Islamic Philosophy', whose principal investigator is Jari Kaukua. It was Kutlu Okan, PhD candidate in that ERC project, who eventually convinced me that I should expand the two articles into one book.

The majority of the book I wrote in the academic year '17–'18, residing at Blackfriars Priory in Cambridge, UK. The home stretch was done as part of my postdoc research grant, sponsored by the NWO (Netherlands Council for Research). I am tremendously thankful for the patronage of the Department of Philosophy and Religious Studies and the Open Access Fund, both at Utrecht University. They made it possible to publish this book in electronic open access. During these years, countless colleagues have helped me along the way, for which I am very grateful.

Rotterdam, Pentecost 2019

Introduction

This book is for humanities students or scholars who are classically trained in handling manuscript materials and wish to take advantage of the incredible computing power at their fingertips but are at a loss where to begin. Some of the more technical parts of the book could be challenging, but a little persistence and practice, over time, should more than suffice.

I also hope to reach those who are more skeptical; who would agree with Paul Eggert, a book historian of modern English literature, when he wrote: “Speaking as a humanities scholar who lacks programming skills and ongoing access to a funded computing laboratory, the assumed advantage of the electronic environment is far less clear.”[1] My response to this challenge is twofold. First, as colleagues around you introduce computer-supported solutions into their workflow, they will gain an edge over you. In fact, as digital methods gain wider currency, digitally restyling parts of our workflow will become the norm, and you may well be left behind if you decide not to do likewise. Second, and more importantly, I will argue that using computers for your erstwhile methodology and workflow requires some adaptions. This is chiefly because digital photos impose specific limitations: their resolutions might be very low, their colors might not be true to life, or they can only be accessed from the museum website. Knowing how to spot and judge these limitations will be most useful to work efficiently and accurately, and this requires a little bit of knowledge about what digital photos of manuscripts are.

In my experience, the most daunting aspect of applying computer-supported methodologies in one’s work is its demand for a life-changing choice. Because, in this day and age, although technology is capable of doing more and more, students and seasoned scholars alike have had little or no exposure to it during their training. The result is that the so-called ‘digital humanities’ (DH) pose a real conundrum: either one pretends it does not exist, or one takes it as a specialization at the graduate level. In the second case, one stops becoming a historian of, for example, ancient Greece, medieval Islam, or the long eighteenth century, and is now on the path to becoming something else—a ‘Digital Humanist,’ where the ‘digital’ part dominates over one’s original expertise. As a consequence, such a researcher tends to be restricted to communicating primarily with other ‘Digital Humanists’ and to publishing in their own specialized journals. The first group, meanwhile, does not invest significant time in

[1] Eggert, P. “The Book, the E-Text and the ‘Work-Site.’” pp. 63–82 in *Text Editing, Print and the Digital World*, edited by M. Deegan and K. Sutherland. Surrey: Ashgate, 2009, p. 63.

© CORNELIS VAN LIT, 2020 | DOI:10.1163/9789004400351_002

learning how digital tools can be used to make our work easier and better. In the rush of assignments, teaching duties, administrative burdens, conference preparations, brushing up language skills, and keeping up with developments in the field, it is one task too many to become acquainted with computer-supported solutions. This leads to an almost perfect disconnect between manuscript studies and 'digital humanities.' On the one hand, introductions to manuscript studies seem to overlook the part of our work that happens on computers, assuming that we have access to the material manuscripts.[2] On the other hand, introductions to digital humanities tend to suppose that the researcher already has the text in digital format.[3] The gap between the two is not seriously addressed.[4]

Another major hurdle in engaging with computer-supported solutions is the perceived cost of both time and funds. In digital humanities, research, all too often, is conducted using team-based projects funded by generous grants. The rationale behind this is that expertise in digital humanities can best be partitioned into two groups: those specializing in the 'digital' aspect and those who focus on the 'humanities' aspect. At the simplest level, this would result in teams of two experts, the one being a humanities scholar and the other a technician or a developer. While the scholar would develop the research questions and the conceptual path to a solution, the technician would make it happen. Experience shows, however, that what Snow calls 'the two cultures problem'[5] becomes almost insurmountable. If the scholarly problem becomes

[2] Notably, even a team working specifically on 'digital editing of medieval manuscripts' produced a guide that assumes there are only material manuscripts, see Haltrich, M., E. Kapeller, and J.A. Schön, eds. *From Sheep to Shelf: An Illustrated Guide to Medieval Manuscripts for Students.* DEMM, 2017.

[3] Blackwell's companion, arguably the field's flagship introduction, fails to mention IIIF and does not have a snippet of TEI to show what it is like, see Schreibman, S., R. Siemens, and J. Unsworth, eds. *A New Companion to Digital Humanities.* Oxford: Wiley-Blackwell, 2016.

[4] Notably, the manuscript specialists from Hamburg include a little about digitized manuscripts in Bausi, A. (et al), ed. *Comparative Oriental Manuscript Studies: An Introduction*. Hamburg: COMSt, 2015. Likewise, the DH specialists Kurz and Jannidis devote a short section in their books to images: Jannidis, F., H. Kohle, and M. Rehbein. *Digital Humanities: Eine Einführung*. Stuttgart: J.B. Metzler, 2017; Kurz, S. *Digital Humanities: Grundlagen und Technologien für die Praxis*. Wiesbaden: Springer Vieweg, 2015. Two resources in English that come close but are a bit too specialized to serve as introductions are the book: Andrews, T., and C. Macé, eds. *Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches*. Turnhout: Brepols, 2014; And the journal issue: "The Digital Middle Ages." *Speculum* 92, no. S1 (2017). See also the series of collected volumes *Kodikologie und Paläographie im digitalen Zeitalter* (5 volumes so far).

[5] Snow, C.P. *The Two Cultures and the Scientific Revolution.* Cambridge: Cambridge University Press, 1959.

more intricate and advanced, it becomes more likely that the technician will fail to provide a solution that truly encapsulates the problem. And when the technology becomes more advanced, it becomes more likely that the scholar will fail to understand how it can be improved to meet their requirements. If we scale back our ambitions towards using and modifying the existing technology, it is possible to have scholars operate on their own, as is customary in the humanities. The tools discussed in this book require no generous grant—they are mostly open source; they are free to download, use, and adapt. Such small adaptions can then be shared back to the community, fostering the organic growth of our toolbox.

Finally, another issue for manuscript studies is that when it comes to digitized manuscripts, there is no prior agreed upon conceptual framework, nor an acknowledged, basic skill set. This book addresses this matter by providing a conceptual and practical toolbox. If something does not have a name, we literally cannot speak or think clearly about it,[6] hence working out a conceptual framework is essential. Similarly, without discussing what we can do with digital tools at the beginner and intermediate level, we cannot properly determine the skills we must develop as a normal part of our methodological toolbox.

Two things should be noted. First, this book is not about digitization itself. Parts of this book, such as Chapter Six, discuss work that leads towards selecting artifacts to be digitized, but for the most part, the assumption is that your manuscripts of interest have already been digitized. If you wish to digitize artifacts yourself, or if you are a professional in this regard, you will profit most from Chapters Two and Three.

Second, neither is this book about plain-text analysis of repositories—such as *Index Thomisticus*,[7] *Thesaurus Linguae Graeca*,[8] *Library of Latin Texts*, *al-Maktaba al-shāmila*,[9] *Chinese Text Project*,[10] *Perseus Digital Library*, or *Project Gutenberg*, to name a few of the largest full-text databases from various fields. Such resources are a major advancement in the humanities, capable of yielding

[6] Carroll, L. *Through the Looking-Glass, and What Alice Found There*. London: MacMillan, 1872, pp. 61–64.

[7] This resource is generally considered to be the first 'digital humanities' project. It was started by Fr. Roberto Busa SJ in 1949 with the aim to have the entire corpus of Thomas Aquinas in electronic full-text format, searchable, indexed, and eventually syntactically analyzed.

[8] The TLG has been a powerhouse since the 70's. I do find it odd that its title is written in Latin.

[9] The name literally means "The All-Comprehending Library" (المكتبة الشاملة).

[10] The Chinese name literally means "The Chinese Philosophical Book Digitization Project" (中國哲學書電子化計劃).

fantastic new avenues of research if we exploit them by automated mining, using a distant reading methodology.[11] But whereas digitization itself is presupposed, such plain text processing is a stage beyond the subject topic of this book. If what you have are images of texts, this is the book for you.

This book represents a case study of how the different aspects of digital manuscript studies may be integrated into the work of one person. This means that most chapters draw heavily from examples of my own work in Islamic studies. The non-Islamicists, I hope, will forgive me for my field-specific examples and find that the problems (and solutions) encountered are universal in nature, applicable to any other field involving digitized manuscripts. Similarly, my focus lies on manuscript and archival materials. Ancient texts or inscriptions, I am aware, may also appear on papyrus, stone, pottery, textile, coins, bone, and other materials. However, for reasons of space and my lack of experience with these materials, I do not address the peculiarities of engaging with such texts through a digital surrogate.[12]

The seven chapters between this introduction and the conclusion are split into two parts. The first three cover the conceptual and theoretical framework of thinking about digital surrogates. The next four explain the practical and technical skills.

Chapter One is theory-laden and, as such, comes logically prior to the rest of the book. Nevertheless, it could very well be read after the other conceptual chapters. In it, a framework is developed to examine the relationships between a material manuscript, print publication, and digital document. Here, I introduce the concepts 'manuscript world,' 'print world,' and 'digital world,' and I discuss how our work can be explained through the different relationships between these worlds. The manuscript world is a realm in which participants use and produce texts by writing them with ink by hand, on parchment, or paper. The print world is a world in which texts are machine-printed on mass-produced paper. Last, the digital world comes into existence when you type

[11] This term was popularized by Franco Moretti as an opposite to 'close reading.' It has found diverse adaption in different fields of the humanities. For my own interpretation, see Lit, L.W.C. van. "Commentary and Commentary Tradition: The Basic Terms for Understanding Islamic Intellectual History." *MIDEO* 32 (2017): 3–26. Lit, L.W.C. van. *The World of Image in Islamic Philosophy: Ibn Sīnā, Suhrawardī, Shahrazūrī, and Beyond*. Edinburgh: Edinburgh University Press, 2017.

[12] That is not to say that there is no interesting work done on them. See e.g. the successful markup standard for epigraphy called EpiDoc: Bodard, G. "EpiDoc: Epigraphic Documents in XML for Publication and Interchange." pp. 101–118 in *Latin on Stone: Epigraphic Research and Electronic Archives*, edited by F. Feraudi-Gruénais. Lanham: Lexington Books, 2010.

on a computer keyboard and see your input appear on an electronic screen. When we edit, we base our work on artifacts from the manuscript world. We work, meanwhile, on a computer, that is, in the digital world. Our final product, however, is often times a printed book, part of the print world.

Chapter Two forms the book's conceptual core. I discuss the perception scholars have of digitized manuscripts as 'larger-than-life' objects, emphasizing the ability to zoom in and make the tiny details invisible to the naked eye visible. I discuss how this perception rests on larger trends of thinking about mechanical reproduction and digital surrogates, in effect viewing digitized manuscripts as though they are a window through which one can look at the material manuscript. As such, it is unsurprising that scholars cite the material manuscript when, in reality, they make use of a digital surrogate. I also respond to the opposite view; that digitized manuscripts destroy the pure experience of handling the material manuscript itself. I argue that both views result from ignoring the 'digital materiality' of digital photos, which in turn occurs because we do not have a vocabulary to describe its features. I propose ten aspects by which to evaluate a digitized manuscript and its repository: (1) size of the collection; (2) online availability; (3) ability to download; (4) the portal through which the repository is accessed; (5) the viewer; (6) indication of page numbers; (7) image resolution; (8) color balance; (9) lighting; and (10) how the image is cut.

Chapter Three shows how these ten aspects can be used to evaluate twenty repositories, which are chosen to give a representative picture of the state of digitization of Islamic manuscripts worldwide. Since many of these libraries also host manuscripts of other disciplines, readers from beyond Islamic Studies should still find this informative of the general state of digitization. The result is that quality and usability varies wildly. Not all manuscripts are downloadable, and the legal restrictions applied to them are often ambiguous. I end this chapter by speculating on the future of these repositories.

Chapter Four starts off the practical part of the book with two topics. First, I discuss how manuscript research specifically concerned with computer-supported solutions has evolved into team projects supported by big grants. I wish to highlight, in particular, when such big projects work well and when they seem to fall short of expectations. From this, we learn that we cannot rely on such teams to produce technical solutions to our problems. Instead, we need to take matters in our own hands. As the first step in this direction, I provide a practical example of how a tablet and free drawing software can be used to perform simple yet effective paleographic work. This part of the chapter is an extended and more in-depth version of an article I previously published, which discussed three glyphs that appear in a text by twelfth-century philosopher Suhrawardī, who claims that only the initiate will understand how these symbols represent the essence of his philosophy. By (literally) drawing from several medieval manuscripts and combining different versions of the glyphs, I arrive at the interpretation that the symbols are constructed from Arabic letters.

Chapter Five discusses the workflow of digital editing. The particular software one uses will change over time, but many of the technical standards on which digital documents rest will remain the same. For this reason, it is essential to know these standards, such as Unicode, TEI, and IIIF. To lower the barrier for working in these standards, I shall provide some pointers on how to set up your computer—for example, how to create your own keyboard layout. The chapter finishes with a discussion of the pros and cons of what is called a 'digital edition': a publication that does not appear in print and, therefore, can take on digital forms unimaginable in the print world.

For digital publishing of any kind, knowing web development technology is a terrific asset. In Chapter Six, I explain the entire process of creating an online catalog of a hitherto uncatalogued collection. We first look at how computers and smartphones can be helpful for fieldwork in an archive and then turn to create an interactive website to make the catalog openly available to others, using HTML, CSS, JSON, and JavaScript. We finish by looking at how those same technologies can help in creating attractive and insightful diagrams.

Chapter Seven adopts a notably technical character. I explain how one might use a simple programming language such as Python and a well-known function library called OpenCV to analyze the covers of several thousand digitized manuscripts. The method is automated image recognition, and it aims to say meaningful things about the shape of the codex. All the while, the core skills for programming that are explained can be applied to any other use case. As such, the chapter revolves around introducing programming in general and Python in particular.

Chapter Eight, the conclusion, is divided into three sections. The first stresses the point that 'digital' humanities and 'classical' humanities are not contradictions; that the latter will benefit from working in the digital world. There has, for this reason, never been a better time to conduct classical philological, codicological, or paleographical studies. The second section synthesizes the lessons learned from the book. In the third section, I offer my perspective on the future of the ongoing incorporation of digital assets and tools within humanities.

This book has two additional parts. As a postscript, I include a few stories about my experience in handling digitized manuscripts in a style similar to

Ignaz Kratchkovsky's memoir *Among Arabic Manuscripts*.[13] With these stories I wish to show that the experience of reading a material manuscript may be destroyed by using a digital surrogate, but other experiences of equally personal and emotional quality come about. It also gives additional insight into my daily practice concerning obtaining, storing, and retrieving digitized manuscripts.

The last part of this book is not actually in this volume, but is a companion website, a digital appendix. You can access it through the URL right below. There you will find images, code, data and other relevant digital documents. In this book itself you will find no URLs or DOIs: they have all been moved to the digital appendix. In fact, many more online resources, technologies, and tools are listed there to give you the opportunity to explore specialized or more advanced options after you have acquired the foundation that this book offers.

For the digital appendix go to GitHub.com/Among

For the stable, citable repository go to Zenodo.org/record/3371200 and use this DOI: DOI 10.5281/zenodo.3371200

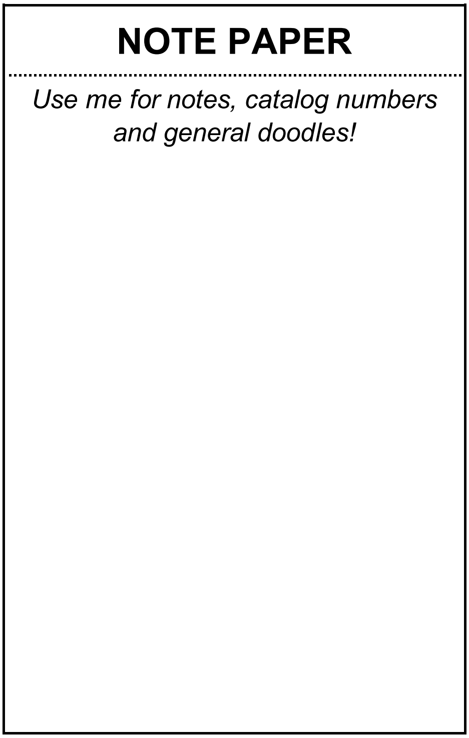
Additionally, you will find a QR-code for every chapter. It will take you to the correct folder within the digital appendix.

[13] Kratchkovsky, I.Y. *Among Arabic Manuscripts*. Translated by T. Minorsky. Leiden: Brill, 1953 [Reprinted 2016].

CHAPTER 1

Manuscript World, Print World, Digital World

1 Three Worlds



NOTE PAPER

Use me for notes, catalog numbers and general doodles!

This is how the slip of paper I found at the Cambridge University Library looked like.[1] Its font and layout clearly tell us that it was originally made on a computer: it is a *born-digital document*. By means of mechanical application of ink onto paper, it was transformed into a *print publication*. Its use, however, is primarily for people to scribble (Latin: *scribere*) on by hand (Latin: *manus*): it is a *manuscript* in the making. It exists as all three simultaneously, though few would recognize it as such. Its digital aspect would elude even the most veteran library patrons. Its creator, quite likely, did not think of it as a print publication. And apparently, Cambridge University's current students have a

[1] Note the mischievous absence of the Oxford comma.

hard time figuring out that it is a potential manuscript, given that its function needs to be called attention to.

Such cross-over artifacts populate our world. And digitized manuscripts, the topic of this book, is a prime example of this. In this chapter, I analyze the double or triple natures of such artifacts to understand their implications.

I wish to frame my analysis by arguing that material manuscripts, printed publications, and digital documents each form what I call a *world*, each with its own communication system, its own episteme or consciousness, its own power structure, and its own social dynamic. The reason I chose 'world' and not, for example, 'era' is illustrated already in this example. Manuscripts, printed publications, and digital documents do not live independently of each other, as though they would convey the same message but in a different medium, at a different point in history. Words like 'condition' or 'culture' also do not cover the entire phenomenon,[2] as they can only be understood when contrasted with a notion like 'technology,' as we shall see. It is true that the technology for manuscripts, printed publications, and digital documents, and the cultures they produced, arose in different parts of history and can be arranged in a rough chronology.[3] Nonetheless, they also do exist simultaneously, sometimes even combined in one artifact, as is the case with the library slip.

The oral/written dichotomy, it may be noted, plays no role in this analysis. Our ways of looking at the world rely too much on writing to consider the dichotomy having great explanatory value.[4] One aspect that sheds light on the differences between manuscript, print, and digital world, and is therefore worth mentioning here, is the representation of the authority of orality. In most serious cases, such as the court of law or political committees, people are summoned to appear in person and *tell* what they know, not just *write*.[5] It is

[2] Illustrated nicely by Jerome McGann who uses both and also 'era' and 'age'. McGann, J. *A New Republic of Letters: Memory and Scholarship in the Age of Digital Reproduction*. Cambridge Mass.: Harvard University Press, 2014. Harold Love gives a succinct overview of different takes on this, from which I follow the first take, which looks at 'print culture' as a 'noetic world.' Love, H. "Early Modern Print Culture: Assessing the Models." pp. 45–64 in *Parergon* 20, no. 1 (2003).

[3] Too many studies take chronology as their analytical framework, e.g. Bolter, J.D. *Writing Space: Computers, Hypertext, and the Remediation of Print*. Mahwah: Lawrence Erlbaum, 2001; Eliot, S., and J. Rose. *A Companion to the History of the Book*. Oxford: Blackwell Publishing, 2007; cf. Finkelstein, D., and A. McCleery. *An Introduction to Book History*. London: Routledge, 2005, p. 17; cf. Dagenais, J. *The Ethics of Reading in Manuscript Culture: Glossing the Libro de Buen Amor*. Princeton: Princeton University Press, 1994, p. 16; McGann, pp. 10, 23; Love, p. 46.

[4] Ong, W.J. *Orality and Literacy*. London: Routledge, 2002 [Or. 1982], pp. 77ff.

[5] Cf. Ong, p. 94; Love, p. 51; Pedersen, J. *The Arabic Book*. Translated by G. French. Princeton: Princeton University Press, 1984, p. 17.

striking how the manuscript world takes over this authoritativeness: a contract still needs to be hand signed in most cases, even if it is only going to be scanned and stored digitally. Similarly, an autographed copy of a printed book can be worth a lot more than a regular copy. In other words, some of the print's value is derived from bringing it into the manuscript world.

My starting point, rather, is that matter matters;[6] the same text, say, the Koran or Homer's Iliad, is fundamentally different when conveyed through a manuscript, a print edition, or a digital version. Analyses of all three distinct worlds are very rare.[7] Scholars often conflate them, either grouping manuscripts and printed publications together against digital documents, or considering printed publications and digital documents as the same against manuscripts. All three can even be conflated to the same thing, arguing that abstract ideas are neutrally conveyed through a medium of one's choice, as an idea can stand on its own, without an expression, while an expression is meaningless without a referent. This perhaps rests on principles in Ancient philosophy, such as the ontologically real precedence of form over matter, active over passive, and intellectual over sensory perception—principles that have loomed large until the Enlightenment, and even then it found some home within Idealism. That, however, would be a discussion about ontology and epistemology in a vacuum.[8] Any person comes about in a social context. Consciousness itself, in the sense of one's 'inner dialogue,' can arguably only arise through the signs one's culture—especially language—provides.[9] But language, in turn, is decided upon previously and, therefore, is hardly a neutral medium. It is, rather,

[6] Cf. Kittler, F.A. *Gramophone, Film, Typewriter*. Translated by G. Winthrop-Young and M. Wutz. Stanford: Stanford University Press, 1999 [Or. 1986], p. xxxix; Hayles, N.K. *Writing Machines*. Cambridge Mass.: The MIT Press, 2002, p. 6.

[7] Notable exceptions are Johnston and Van Dussen who speak of "manuscript culture," "print culture," and "digital culture." Johnston, M., and M. Van Dussen. "Introduction: Manuscripts and Cultural History." pp. 1–16 in *The Medieval Manuscript Book: Cultural Approaches*. Cambridge: Cambridge University Press, 2015.

[8] Like Descartes attempted with his *cogito*-argument, or Ibn Tufayl attempted with his allegory of the man growing up alone on an island.

[9] Personally, I think that the *mode of expression*, the garb under which real knowledge expresses itself in the mind, is indeed subject to contextualization. I would call this a localized truth. However, this does not preclude a belief in transcendental, non-discursive thoughts that can nonetheless be called knowledge, real knowledge, objective and invulnerable to localization. As McLuhan writes, "The content of writing is speech […] If it is asked 'What is the content of speech?,' it is necessary to say, 'It is an actual process of thought, which is in itself nonverbal.'" McLuhan, M. *Understanding Media: The Extensions of Man*. Cambridge Mass.: MIT Press, 1994 [Or. 1964], p. 8. Cf. Wittgenstein's famous statement "The limits of my language mean the limits of my world." Wittgenstein, L. *Tractatus Logico-Philosophicus*. Translated by D.F. Pears and B.F. McGuinness. London: Routledge, 2001 [1921], p. 68.

an “ideological phenomenon par excellence,” as Valentin Volosinov puts it, concluding that “the individual consciousness is a social-ideological fact,”[10] which is best revealed “in the material of the word.”[11] Indeed, not only is it best revealed, but it is also only able to come about by its materiality. Signs that fall out of use lose their force, “becoming the object not of live social intelligibility but of philological comprehension.”[12] Material signs, in that sense, have a worldview-making potential that needs to be re-activated constantly in others to keep a sense of community fresh in our imagination,[13] and indeed to preserve its own power to which people have submitted themselves.[14] Otherwise, a process of disenchantment (*Entzauberung*), hard to reverse, occurs. Seen this way, Nelson Goodman can turn that Idealist paradigm upside down. Instead of ideas being able to do without expressions, he says:[15]

> Although conception without perception is merely *empty*, perception without conception is *blind* (totally inoperative). Predicates, pictures, other labels, schemata, survive want of application, but content vanishes without form. We can have words without a world but no world without words or other symbols.

The word ‘form’ here has flipped its meaning. Whereas in Ancient philosophy it is used to denote the essence that is abstract from matter, for Goodman it indicates the specificity of exactly that matter, which is best specified through words. And, as Ong has demonstrated, those words are written words. In “the new world of writing,” he says, we are “beings whose thought processes do not grow out of simply natural powers but out of these powers as structured, directly or indirectly, by the technology of writing.”[16] Such an ontology, based not on substance but function,[17] constructs the world as a social world, this

[10] Volosinov, V.N. *Marxism and the Philosophy of Language*. Translated by L. Matejka and I.R. Titunik. New York: Seminar Press, 1973, p. 12.

[11] Volosinov, p. 14.

[12] Volosinov, p. 23.

[13] Anderson, B. *Imagined Communities*. 2nd ed. London: Verso, 2006.

[14] Althusser, L. “Ideology and Ideological State Apparatuses.” pp. 127–88 in *Lenin and Philosophy and Other Essays*, translated by B. Brewster. New York: Monthly Review Press, 1971, p. 132. Cf. also Levi-Strauss’s remark on writing as facilitator of (societal) slavery: Lévi-Strauss, C. *Tristes Tropiques*. Translated by J. Russell. New York: Criterion Books, 1961, pp. 291–293.

[15] Goodman, N. *Ways of Worldmaking*. Indianapolis: Hackett Publishing, 1978, p. 6.

[16] Ong, p. 77.

[17] Goodman, p. 7.

construction occurring through the continuous re-activation of signs.[18] Such continuous re-activation by one person can be called a habit, and a recurring habit across society can be called an institution.[19] Such institutions are built from people's behavior, but if they grow big enough, the power dynamic reverses and institutions stipulate people's behavior, thereby giving boundaries to their ethics. This is especially true for later generations of people who consider institutions as a matter of fact rather than a social construct. As Peter Berger says, "the reflective integration of discrete institutional processes reaches its ultimate fulfillment. A whole world is created,"[20] and so we see the confirmation for our choice of the term 'world' again. Furthermore, this world "is experienced as an objective reality."[21] If Berger is right, it means that the moment this worldview is created, in a dialectical, hermeneutic circle between the individuals and the institutions, the next area of philosophy to be influenced is epistemology. The signs that are given to an individual to express themselves provide the framework in which new items of knowledge are to be expressed. In this sense, it provides boundaries to the knowable,[22] defining what a good question is and what a good answer might look like. This, it seems to me, is relatable to Michel Foucault's notion of *episteme*, which "defines the conditions of possibility of all knowledge."[23] To get grips of such an episteme, he asks us to look at "surfaces of their emergence," "the authorities of delimitation," and "the grids of specification."[24] These can be related respectively to the individuals, the power of institutions over habits, and the influence of habits on institutions. I bring up Foucault's version of this phenomenon not only to provide a way in for those familiar with his work, but also because his three-point description is useful for our context. In this book, we set up this theoretical framework in order to understand how material manuscripts, printed publications, and digital documents each form a world, how these worlds are different from each other, and what happens when these world intermesh (or collide).

On a most general and literal level, the 'surfaces of emergence' are, of course, the vellum of the manuscript, the page of the printed publication, and

[18] Cf. Dagenais, p. 14.

[19] Berger, P.L., and T. Luckmann. *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. London: Penguin Books, 1966 [Reprint 1991], pp. 70ff.

[20] Berger, p. 114.

[21] Berger, p. 77.

[22] Berger, p. 83.

[23] Foucault, M. *The Order of Things: An Archeology of the Human Sciences*. New York: Vintage Books, 1994 [Or. 1966], p. 168.

[24] Foucault, M. *The Archeology of Knowledge*. Translated by A.M.S. Smith. New York: Pantheon Books, 1972, pp. 41–42.

the electronic interface of the digital document. These 'surfaces,' I suggest, can be further split between society-based surfaces and object-based surfaces. For society, we shall consider the notions of archive, time, and the dichotomy of center/periphery. For objects, we shall consider the notions of ecosystem, page, and transparency.

As institutions, 'authorities of delimitation,' we can distinguish three levels. One metaphysical, worldview level, one at the level of society, and one at the level of individual actors. The first one consists of the notions of reality and truth. The second consists of canon. I have formed the last using the notions of gatekeepers and producers.

For the 'grids of specification,' the habits, I wish to structure the analysis by three subcategories: possession, engagement, and product. For possession, we shall use the notions of monetized product, ownership, and value. For engagement, we have closely related terms of activity, conversation, reading, and response. For product, we use the illusion of fixity, and product.

All of these notions make the manuscript, print, and digital worlds remarkably different from each other. My analysis proposes the trichotomies as sketched out in the tables below. The differences were condensed into as few words as possible, preferably one, to make the summary overview as compact as possible while maximally highlighting the differences. For some of the trichotomies, you might easily understand what I am pointing out; nonetheless, the next half of the chapter is devoted to a more in-depth description of each notion. If you rather prefer to see these differences put into practice, you can go over to the second half of this chapter, which consists of three case studies that use these notions to highlight the difference between manuscript and print world, print and digital world, and manuscript, print, and digital world, respectively.

TABLE 1.1 Surfaces of emergence

	Manuscript world	**Print world**	**Digital world**
Page	arbitrary	neutral and fixed	non-existent
Transparency	little	some	greatest
Ecosystem	faint notion	agnostic	acknowledged
Center	local	regional	global
Archive	just-as-requested	just-in-case	just-in-time
Time	enduring	progressive	of the moment

TABLE 1.2 Authorities of delimitation

	Manuscript world	**Print world**	**Digital world**
Producers	reader	medium	author
Gatekeepers	society	publishers	tech industry
Canon	continuous	staggered	haphazard
Reality	reality-making	reality-stating	reality-replacing
Truth	pre-truth	one truth	post-truth

TABLE 1.3 Grids of specification

	Manuscript world	**Print world**	**Digital world**
Monetization	codex	text	user
Value	a lot	some	a little
Ownership	privately owned, privately made	privately owned, corporately made	corporately owned, corporately made
Activity	active	reactive	interactive
Reading	charitable	docile	egotistical
Response	in the margin	another text	within the text
Conversation	dialogue	monologue	soliloquy
Product	process in flux	end result	end result in flux
Fixity	non-existent	strong	illusion

1.1 *Surfaces of Emergence*

1.1.1 Object-based

The sequence of words that forms a text can manifest itself in any of the worlds, manuscript, print, and digital. In the manuscript world, its fundamental surface of emergence is the folio. This can be of any material, but it is most likely papyrus, parchment, or paper. Its size and color are largely inconsequential; variations in them are taken as a matter of fact and whatever is needed to complete copying the text is used. Further, the entire folio is of importance, in contrast to print publications, where, in principle, only the text block is important. For manuscripts, however, as John Dagenais famously noted, “it is at the edges of manuscripts […] that the most important part of ‘medieval literature’ happens.”[25] For printed publications, the **page** is the surface of emergence.

[25] Dagenais, p. xvi.

This is a carefully chosen delimiter of the text.[26] Each page is meant to be cut alike and is often white as snow. It bears with it a great consequence, namely the page number, which functions in the print world as the ultimate neutral decider of the place of a text. For example, at the final stage of print proofs, authors are asked not to request changes that would push text onto the next page. Page numbers have assumed such great power that their functionality has been projected onto the manuscript and the digital world when referring to a source from either world. For the digital world, the fundamental surface of emergence is the interface, by which I mean more than just the pixels in a monitor. The interface is the entirety of the user experience, from storage in bits and bytes, through a file type and operating system, to a monitor and peripherals to give instructions to the computer. The durability of these surfaces is notably different; manuscripts, especially those of parchment or paper, can be incredibly durable and sturdy, easily surviving many centuries. Print publications have often greatly reduced the quality of their materials. With wood pulp mixed in paper and using glue for binding instead of sewing the quires, modern books can quite literally fall apart into unusable pieces. Digital documents, in turn, have seen a very fast outdating. Whereas hardware upgrading usually only means improvement, software upgrading can render file types obsolete and unusable. Within mere decades, a digital document can become practically inaccessible.

Despite the volatility of digital documents, the **transparency** of its surface of emergence is the greatest of all.[27] This is encouraged by tech companies which hide as much as possible from the user. Jason Merkoski, former lead-engineer of Amazon's electronic reading device says: "The Kindle itself is just the tip of the iceberg, and its true workings are invisible. That's exactly how Amazon wanted it to be."[28] The fact that somebody accesses the text through a specific hardware and software setup remains virtually always unacknowledged. Rather, as Lori Emerson has analyzed well, the interface of digital documents is purposely made as transparent as possible using a blinding seduction of offering supposed natural usability without the need of a manual.[29] She concludes

[26] Mak, B. *How the Page Matters*. Toronto: University of Toronto Press, 2011, esp. p. 3, 73.

[27] Kichuk, D. "Metamorphosis: Remediation in Early English Books Online (EEBO)." pp. 291–303 in *Literary and Linguistic Computing* 22, no. 3 (2007), pp. 296–297; Foys, M.K. "Medieval Manuscripts: Media Archaeology and the Digital Incunable." pp. 119–39 in *The Medieval Manuscript Book: Cultural Approaches*, edited by M. Van Dussen and M. Johnston, Cambridge: Cambridge University Press, 2015, p. 119.

[28] Merkoski, J. *Burning the Page: The Ebook Revolution and the Future of Reading*. Naperville: Sourcebooks, 2013, p. xvi.

[29] Emerson, L. *Reading Writing Interfaces: From the Digital to the Bookbound*. Minneapolis: University of Minnesota Press, 2014, esp. pp. x, 2–5, 24.

that this veil of ignorance can, does, and will change the nature of the digital world from one of true creativity towards merely being offered predetermined choices that are carefully restricted by corporations. She concludes that "now, digital interfaces are artful only to the extent that they don't work,"[30] as only when digital interfaces do not work is the veil removed and we are reminded that there must be inner workings to the technology we use.

In the limited cases that digital documents are considered in their materiality, users pretend they are (like) print publications.[31] A computer still provides, as it were, a transparent window onto the document.[32] McLuhan already wrote in the sixties that "the electric light escapes attention as a communication medium just because it has no 'content.'"[33] This transparency is augmented by the shared control of writer and reader over the text's appearance (in fact, both think they have sole control, not considering the other).[34] These are fundamental misunderstandings, though easy to make within the paradigm of the digital world. In this paradigm, the vocabulary is concentrated upon user experience, with the computer merely seen as a tool that can be used in whatever way the user wants to. The medium itself is transparent, invisible, and its influence on the message is not considered. It is this mistaken transparency that is the subject of the next chapter, where we discuss the phenomenon of digitized manuscripts. For print, transparency means the page and the uniformity of the print faces.[35] In its supposed neutrality, it fades into the background for the reader. This is how Stephen Nichols remarks that "the medieval artifact, for Spitzer, was the edited text,"[36] not the manuscript evidence which underlies the critical edition. Not so for manuscripts, where readers are much more aware of the specificity of the folio, its ornamentation, even the style

[30] Emerson, p. 4, cf. her discussion of easter eggs and glitches, pp. 24, 36.

[31] Mandell, L. *Breaking the Book: Print Humanities in the Digital Age*. Chichester: Wiley-Blackwell, 2015, p. 3; Sutherland, K. "Being Critical: Paper-Based Editing and the Digital Environment." pp. 13–26 in *Text Editing, Print and the Digital World*, edited by M. Deegan and K. Sutherland. Surrey: Ashgate, 2009, p. 19; Eggert, P. "The Book, the E-Text and the 'Work-Site.'" pp. 63–82 in *Text Editing, Print and the Digital World*, edited by M. Deegan and K. Sutherland, Surrey: Ashgate, 2009, p. 67; Weiss, A. *Using Massive Digital Libraries*. Chicago: ALA TechSource, 2014, pp. 7–8.

[32] Cf. Sutherland, "Being Critical," pp. 17–18.

[33] McLuhan, p. 9.

[34] Lanham, R.A. "The Electronic Word: Literary Study and the Digital Revolution." pp. 265–290 in *New Literary History* 20, no. 2 (1989), p. 266.

[35] Lanham, p. 266; Mak, p. 8.

[36] Nichols, S.G. "Introduction: Philology in a Manuscript Culture." pp. 1–10 in *Speculum* 65, no. 1 (1990), p. 3.

of handwriting. Perhaps only the letters themselves provide some sense of transparency—as a neutral, effortless means to evoke a certain letter, sound, or word in the reader's mind.

We can also look at the relationship of one object to another: their **ecosystem.** For manuscripts, each codex is individuated with only a faint notion of it being a copy or exemplar of another manuscript containing the same text. Meanwhile, for print publications, each copy is only a copy. Readers, then, as far away as they may be from each other, share the space that the page dictates since each copy has the same space on a specific page. However, readers are largely agnostic to this as they cannot directly see that they share the space, for example, by interaction. This would be possible for digital documents, which can truly allow for sharing space by enabling multiple users to use one file. Prime examples include the comments section under a news item or a Wikipedia entry.

1.1.2 Society-based

Societal surfaces of emergence are informative as well. Take for instance the **center/periphery** dichotomy. For manuscripts, local centers are possible as the periphery can produce unimpeded by the power of the center. In the print world, the periphery, one might think, has the leverage to burst out to the rest of the area as one could at once produce and disseminate as many copies as one would like. This, however, is but an illusion: the means of production are too costly to be operated by anybody. The center takes a firmer grip on the periphery. In the digital world, the discourse takes place in its hardware-location, which means that there can be a global center.

Of a different nature is the notion of the **archive**. For the manuscript world, one codex already is an archive. It can contain different texts, and the margins can be filled with readers' comments. In the world of print, we have a veritable archive fever. This is because the printed paper itself is authoritative, and so it is almost as though whoever has the most paper is most powerful and most correct. There is, therefore, a great incentive to collect print publications and combine them into shelves and shelves of archives. For digital materials, it is not so much the collecting that is of importance. It is important not to be reliant on others, but the collection itself can happen quick and cheap. It is rather the structure and searchability of the archive that is important. If manuscripts are acquired 'just-as-requested' as a self-containing unit which has space within itself for growth, either by sewing in additional quires or writing in the margins, then print publications are acquired 'just-in-case,' as their authority will be of weight whether it is ever touched or will forever sit in the archive.

And in that case, digital documents can be said to be obtained 'just-in-time,' with much fewer demands on the shape and form as long as it can be found in a reasonable time (which, for the digital world, can sometimes be less than a minute).

This brings us to the aspect of **time** on the surfaces of emergence. Manuscripts, by connecting people who are centuries apart, transcend time. A manuscript can lie dormant only to be picked up three-hundred years later, at which a reader can write notes in the margin, talking back to the author as though they were a contemporary. Print publications, however, have a progressing aspect to them. Whatever is published is the new authority, superseding previous publications. Since the publication is made available to the entire nation or linguistic zone, there is a sense of everybody being lifted onto the next plane of knowledge. The progression is also visible in intertextuality, where the most often cited texts are those that have recently been published, neither immediately nor far in the past. The digital world, in turn, operates in 'real-time.' Especially those documents connected to the internet work best when they reference events that have just happened, or are happening right now. Since documents can be updated without any trace of such operation, a sense of progression is lost.

1.2 *Authorities of Delimitation*

1.2.1 Actors

Let us start with the most basic and directly active authorities, what I have called actors. In terms of the **producer** of a text, the difference between the three worlds could not be greater. For manuscripts, the reader is the producer.[37] For print publications, the medium—that is, the system of different businesses involved in manufacturing and selling books—is the producer.[38] For digital documents, it is the author who is the producer.

This is further explained by looking at the different purposes of production. If the above is true then the purpose of manuscripts is to be read, the purpose of print publications is to be transmitted, and the purpose of a digital document is to be written. When a reader in a print world wants to read something, they simply go to a store and buy a copy. However, this only has a bearing on the production of that copy in a second degree; the copy was already made.

[37] Cf. Johnston and Van Dussen, p. 5; Dagenais, pp. xvii, pp. 23–24; Blair, A. "Afterword: Rethinking Western Printing with Chinese Comparisons." pp. 349–361 in *Knowledge and Text Production in an Age of Print: China, 900–1400*, edited by L. Chia and H. De Weerdt, Leiden: Brill, 2011, p. 353.

[38] I strongly disagree with Walter Ong's assertion that "manuscript culture is producer-oriented" and "print is consumer-oriented." Cf. Ong, p. 120.

Compare this with the manuscript world, in which copies of texts are seldomly for sale. Thus, as a rule, if somebody wants to read a text, they need to acquire a copy by literally copying it out from another copy, in other words, the production of that copy is specifically for that reader.[39] Whether readers do that themselves or ask professional scribes to do it is not important, as the primary agency of the reader is the same in both cases; the reader sets the process of production in motion out of a desire to read.[40] As Johannes Pedersen describes, even the first time a text is written by its author, it is done in a setting requiring readers, listeners, and a process of back-and-forth proposing and approving.[41] Not so in the print world. In a span of a few years, in the 1510s, it is said that over 300,000 copies of Luther's tracts were produced.[42] Could they all have found a customer? And even so, has each copy been read? The answer is, of course, no. As the bibliographer Hugh Amory pithily said: "most printed books have never been read."[43] In the same vain, when the early 16th-century emperor Maximilian I introduced print in his chancellery, he ordered draft versions to no longer be crossed out or erased but stored.[44] Having shelves and shelves of print material became in itself a statement of power and truth. Even today, dissertations are written only to be printed and placed on a library shelf, its only effective power, throughout the years, reduced to bend the shelf ever so slightly.[45] In this sense, it is true that for print, the medium is the message.[46]

In the digital world, transmitting a text is no effort, and because of the many components required, the medium itself is hardly fixed. Indeed, the digital world has moved only more towards interoperability, where the same file may be presented differently based on the device it is opened on. The one bringing a digital document about is, then, the writer. In fact, in the digital world, we can better speak of users, persons who are readers and writers simultaneously. With the provided interactivity, digital documents hardly invite you to

[39] Bourgain, P. "The Circulation of Texts in Manuscript Culture." pp. 140–159 in *The Medieval Manuscript Book: Cultural Approaches*, edited by M. Van Dussen and M. Johnston, Cambridge: Cambridge University Press, 2015, p. 147.

[40] Cf. Dagenais, p. xviii.

[41] Pedersen, pp. 31, 34.

[42] Finkelstein, p. 52.

[43] Hugh Amory, as cited by Blair, p. 352.

[44] Vismann, C. *Files*. Translated by G. Winthrop-Young. Stanford: Stanford University Press, 2008, p. 92.

[45] Dunleavy, P. *Authoring a PhD: How to Plan, Draft, Write and Finish a Doctoral Thesis or Dissertation*. New York: Palgrave Macmillan, 2003, p. 274.

[46] McLuhan, pp. 7–21. As John Dagenais rightly points out for the manuscript world: "It is not so much the medium that is the message as it is the process of production." Dagenais, p. 18.

be a reader merely. For example, people seldom derive joy from merely 'liking' other people's statuses on social media. Rather, people are usually motivated to post a status of their own and receive likes from others.[47] This episteme places stress on notions of curation and erudition since it is tempting, from a digital point of view, to merely offer the evidence (all the evidence) and let the user decide for themselves what to do with.

The different emphasis on reading, medium, or writing also explains how the power dynamic of text production is different in each world, that is, who the **gatekeepers** are. In the manuscript world, it is society as a whole that decides the allowable and the not-allowable by the cumulative effect of copying or neglecting to copy texts. This makes for a decentralized form of authority.[48] Not so for the print world, where, it seems, everything is held together by commercial forces. When we consider the models of (printed) book history by Darnton and by Adams and Barker, we notice how both allow a connection between author and reader through a closed system, by the mediation of publishers, printers, shippers, and booksellers.[49] The mere fact that they could draw up models like this shows the integrated nature of book publishing in the print world. When a similar model would be drawn up of the manuscript world or digital world, it would be a lot more busy and not as circular as it is for the print world. For, there is in the digital world little standing in the way for people to write and publish. Gatekeepers in this world are the software companies who set the technical limitations of what is possible but they are in this sense much further away from the actual production of texts than publishers are in the print world.

1.2.2 Society

These actors together also bring about institutions of authority at the societal level, irreducible to one or the other actor but nonetheless influential.

The authority of a **canon** is noticeably different among the worlds. In the manuscript world, a canon builds up slowly but democratically, as each voice can be barely controlled. A manuscript canon comes close to a Darwinian

[47] A particularly good example is the rise of fan-fiction in the digital world. See Lindgren Leavenworth, M. "Paratextual Navigation as a Research Method: Fan Fiction Archives and Reader Instructions." pp. 51–71 in *Research Methods for Reading Digital Data in the Digital Humanities*, edited by G. Griffin and M. Hayler. Edinburgh: Edinburgh University Press, 2016.

[48] Johnston and Van Dussen, pp. 9–12.

[49] Darnton, R. "What Is the History of Books?" pp. 65–83 in *Daedalus* 111, no. 3 (1982); Adams, T.R., and N. Barker. "A New Model for the Study of the Book." pp. 5–44 in *A Potencie of Life: Books in Society*, edited by N. Barker. London: British Library, 1993.

system of natural selection; those texts deemed valuable enough are copied by readers, while those which are not are not copied and invariably go extinct. How popular a text is can literally be measured by its 'population,' the number of copies in which it survives.[50] What the print world inherits is the slow rate at which canon is made and developed. This is not exactly because readers take a long time to sort out classic from rubbish, but more so because the production of a print publication is a time consuming endeavor. Print publications are an all-or-nothing deal; once it is committed to print, there is little anyone can do. The process of writing, editing, correcting, and finally typesetting is, therefore, a deliberately drawn out process. Once everything is set to go, the publisher can 'game the system of natural selection' by all of a sudden putting out a population of any number of copies. Since this is solely the publisher's decision, the canon is overtly dictated. Readers' preference can only sway publishers' decisions in a secondary sense, by either leaving a publisher with large stock or buying out the stock so fast that the publisher is moved to make another print run. In all of this, with publishers and the demanding process of publishing as gatekeepers, the voices of authors that make it into the canon are subject to selection.[51] In China, where printing was mostly done with woodblocks, it had an on-demand nature, making the canon produced in that print world much more like a canon in a manuscript world.[52]

Lastly, in the digital world, a canon is formed and changed fast. In fact, the speed of production and dissemination is so fast that often it is not the content that is canonized, but the process. For born-digital, online texts, the process can be a website or a particular writer on a website. For digitized manuscripts, the process can be the digitization technology or the online portal to accessing the photos. Especially in the case of digitized documents, which originally belonged to either the manuscript or the print world, the worth of one artifact can be hard to judge. Its context is missing, both subject-wise and popularity. Whether a book sold a hundred or a hundred thousand copies, both will likely be represented digitally in one document each, and different editions or print-runs of the same book are often overlooked in the digital world.[53] The canon

[50] Wogan-Browne, J., N. Watson, A. Taylor, and R. Evans, eds. *The Idea of the Vernacular: An Anthology of Middle English Literary Theory 1280–1520*. Exeter: University of Exeter Press, 1999, pp. 4–5.

[51] Cf. Mandell, p. 91.

[52] Chia, L., and H. De Weerdt. "Introduction." pp. 1–32 in *Knowledge and Text Production in an Age of Print: China, 900–1400*, edited by L. Chia and H. De Weerdt. Leiden: Brill, 2011, pp. 14–15; Blair, pp. 351, 355.

[53] Patten, E., and J. McElligot, eds. *The Perils of Print Culture: Book, Print and Publishing History in Theory and Practice*. London: Palgrave Macmillan, 2014, p. 13; Weiss, p. 91.

is subtly dictated, indicating that the voices are seemingly free but actually are within processes of control. For example, terms and conditions may suspend users from participation, and algorithms can decide whether to show or suppress the content.

1.2.3 Worldview

We come to the highest, most abstract level of authorities of delimitation—reality and truth—which operate on the worldview level. In effect, these notions are consequences of the previous authorities but are worth mentioning separately because of their saying power. In terms of **reality**, the difference between the worlds is conjunctive with Baudrillard's division of "a universe of natural laws to a universe of force and tensions of force, today to a universe of structures and binary oppositions."[54] We may say that a manuscript is reality-making, a printed publication is reality-stating, and a digital document is reality-replacing (i.e., hyperreality). In a manuscript world, the worldmaking potential is bound up with readers, who need to decide what to read. In order to read, they need to copy a text. It is the physical manifestation of ink on paper that makes the reality of the argumentations of the texts, the culture that these texts constitute and shape. With every act of copying, the reality of this culture, this worldview, is reified. In a print world, it is also the act of printing the black ink on the white paper that reifies the culture, but, as noted before, it is not really up to the readers to decide what is committed to print. Rather, the gatekeepers are publishers, standing above society at large, dictating to it what ought to be considered part of the culture and what not. In the digital world, it is not the commitment of ink on paper that reifies a worldview. In its aspect of reality-making, the digital world largely relies on encoding the "real world" (manuscript and print world) into zeros and ones. Because the choice is only between zero and one, the digital world's reality is one of either: existing truly or not existing at all, indicating that digital documents are often hailed as 'larger-than-life.' In this way, the digital surrogate somehow becomes a better version of the real world object and can, therefore, supposedly rightfully, outright replace the real world object.[55]

Similarly, for **truth**, the manuscript world can be called a pre-truth era, where truth is organically established through the many permutations of readers copying texts and, thereby, selecting the argumentations that are considered truthful. Therefore, the print world can simply be called a truth era, where

[54] Baudrillard, J. *Simulations*. Translated by P. Foss, P. Patton, and P. Beitchman. Semiotext[e], 1983, p. 103.

[55] Patten, McElligot, p. 11; Dagenais, p. 216.

the modernist ideal—of there being only one truth that may be forcefully affirmed—is in effect. Any other expression is deemed corrupt.[56] Meanwhile, the digital world is a post-truth era, the truth not being construed by what you receive but more so by what you tell others. As Baudrillard cites Nietzsche: "Down with all hypotheses that have allowed the belief in a true world."[57]

1.3 *Grids of Specification*

Lastly, we have the grids of specification. For our purpose, we may subdivide the relevant notions under the headers of possession, engagement, and product.

1.3.1 Possession

Under the header of possession, we can discuss what the **monetized product** is for each world. As may be expected, this has a close relation with the stakeholders. In the manuscript world, the product that can be profitably exploited is the codex, the physical object in which the text, any text, is contained. A codex is typically produced with such high quality as to outlive its first user.[58] For print, it is the text itself. Codices can be mass produced and, therefore, profit comes from selling an ever larger number of copies. The text itself is a scarce good in a print world, which is why the publishing industry has almost always lobbied for copyright laws. The text as the monetized product for the print world can be gleaned from Walter Ong's view, who argues that "alphabet letterpress printing, in which each letter was cast on a separate piece of metal, or type, marked a psychological breakthrough of the first order. It embedded the word itself deeply in the manufacturing process and made it into a kind of commodity."[59] The *mise-en-page*, however, is important too, unlike with digital documents.[60] For example, this allows for the page number to become a dominant feature of print publications, from which derives the possibility to reliably and accurately refer to the text.[61] In the digital world, it is yet another thing altogether—the monetized product is the user itself. In abundance and easy to reproduce are both textual content and textual manifestation, whereas in scarce supply is the user engagement.

[56] Ong, p. 128, Mahdi, M. "From the Manuscript Age to the Age of Printed Books." pp. 127–142 in *The History of the Book in the Middle East*, edited by G. Roper. Surrey: Ashgate, 2013, p. 131.

[57] Baudrillard, p. 115.

[58] Johnston and Van Dussen, p. 7.

[59] Ong, p. 116. Cf. Ong, p. 129.

[60] Nunberg, G. "The Places of Books in the Age of Electronic Reproduction." pp. 13–37 in *Representations* 42 (1993), p. 18.

[61] Sutherland, "Being Critical," p. 22.

Inherent to this is the **value** of each individual artifact. In a manuscript world, an artifact is worth a lot, as it itself is the exploitable, scarce good. In the print world, a single artifact represents some value, as its sale is still what gives the publishers profit, be it only the aggregate sale of many copies. Parallel to this runs the intellectual value, the difference of which John Dagenais has aptly described when he asked the rhetorical question "What is the intellectual value (and cultural significance) of taking a text that was written and read in a variety of forms in numerous medieval manuscripts and transforming it into a single printed book?"[62] I use 'parallel' because I think we can equally ask for the monetary value in exchanging one for the other, as is done through critical editing, and the rhetorical value would be retained. There is an increasing trivialization of the written word, from oral to manuscript to print to digital, because, as Jan-Dirk Müller aptly notes: "Technical reproduction might have seemed at first a chance to preserve writing for everyone and for all time, but it abolished traditional selection mechanisms that established what is worth knowing and preserving."[63]

In the digital world, a single artifact represents consequently very little value. Without engagement, its existence is quite meaningless as far as the digital world is concerned. Engagement can only be driven, as BnF's former director Jean-Noël Jeanneney forcefully points out, if a repository is curated: "An indeterminate, disorganized, unclassified, uninventoried profusion is of little interest,"[64] he says. As the digital world and its episteme move forward, Jeanneney notes that we are already up against it: "The enemy is clear," he says, "massive amounts of disorganized information."[65] This perspective is refreshing, since it seems that digital platforms are often built on the premise that 'if we build, they will come.'[66]

To complete the grids of specification concerning the possession of artifacts, we can look at **ownership** from the points of view of the producer and consumer. A manuscript is privately owned and privately made; a print publication is privately owned but corporately made. A digital document, finally,

[62] Dagenais, p. xvi.

[63] Müller, J.D. "The Body of the Book: The Media Transition from Manuscript to Print." pp. 32–44 in *Materialities of Communication*, edited by H.U. Gumbrecht and K.L. Pfeiffer, Stanford: Stanford University Press, 1994, p. 33.

[64] Jeanneney, J.-N. *Google and the Myth of Universal Knowledge: A View from Europe*. Translated by T.L. Fagan. Chicago: The University of Chicago Press, 2007, p. 5.

[65] Jeanneney, p. 70.

[66] Andrew Weiss notes, perhaps more wishfully than factually, that libraries are changing from a 'just in case' strategy to a 'just in time' strategy, in which a resource is acquired only after a demand from a patron, Weiss, p. 130.

is in principle corporately owned and corporately made. For manuscripts, this may be rather obvious; even when texts are copied by a professional scribe and even when such is done without prior commissioning, the production still has a private element to it—in the sense that it can inherently be made according to specific demands, and this would not at all be out of the ordinary. Once the production is completed, the result, the codex, belongs squarely to the readers, who can do with it whatever they want. A new binding, or reshuffling of the text, writing in the margins: none of these codex-altering actions would be considered strange, and that is because it is the prerogative of the owner of the codex to do with it what they want. For print publications, such authority is shared between the reader and the publishing industry, since the ownership can indeed be said to belong to the reader, but the production belongs explicitly and exclusively to the conglomerate of publishers, printers, shippers, and booksellers. From this, we can predict that it could be frowned upon for a reader to do alterations to the artifact. "I don't annotate my books. Personally, I think that defiles the printed page," admits an advocate for digital documents.[67] We may note a process of more and more ownership towards the publishers over the centuries. Whereas before, books could be published in fascicles, and it was normal, expected, of the reader to bind their books to their liking. Today, books are often in paperback format, held together by glue, making it nearly impossible to do anything major to the book without fundamentally breaking it. The digital documents are, in this sense, the logical conclusion of a shift towards corporate ownership. Any digital document only exists and can only be fruitfully used through a combination of hardware and software. Software is more and more provided on a subscription basis, making users rely on corporations to access digital documents. Their user agreements can legally lock anyone out of their own work if users transgress a set of arbitrary rules. The continued development of software and hardware can make a current digital document go out of date in a matter of years with no possibility of further access. Their aggressive push to convince users to physically transfer files from local computers to their corporate servers ('the cloud') means that once those servers go offline, users can no longer access their files. This process is not only happening at the consumer level. With the arrival of large collections of digitized resources in the possession of corporations, such as Brill's *Middle Eastern Manuscripts Online* (combined current cost: $38,370) or Gale's *Eighteenth Century Collections Online* (cost can be more than $300,000,

[67] Merkoski, p. 16.

excluding an annual fee), this ought to be an acute issue for academia across the board.[68]

1.3.2 Engagement

We engage with artifacts in several ways. First, we can consider the kind of **activity** in each world. Manuscripts can be said to be active, in that the medium responds to the user, with the user being both writer and reader, sender and receiver, *écriture* and *ré-écriture*.[69] The medium, the surface of the folios of codices, change according to the actions of writers and readers. For printed publications, the medium only responds to the sender: the author in combination with the publisher. The reader is confronted with the end result and can only react. The print world is, therefore, reactive. In contrast, the digital world is interactive. Although the medium is presented as an end result, users can do all kinds of things with it to change it. In other words, users respond to a medium that responds back. This makes the act of **reading** quite different in each world. To exaggerate the differences, we can say that manuscripts are read charitably and with great engagement and focus. Print publications are read docile and acceptant. And digital documents are read egotistical and only for a specific personal purpose. Often, digital documents are entered not from the beginning of the text or the beginning of a chapter, but at the very instance where a keyword search has brought the reader.[70]

That, in principle, this is so can be seen from the different ways readers can **respond**. In manuscripts, readers can respond by writing in the margins. For printed publications, this is possible but uncommon. Somehow, it seems that the ultra-white of the pages and the crispness of the printed black text resists handwritten comments. Walter Ong observed:[71]

> There is no way directly to refute a text. After absolutely total and devastating refutation, it says exactly the same thing as before. This is one reason why 'the book says' is popularly tantamount to 'it is true.'

[68] Patten, McElligott, p. 14.

[69] Dagenais, p. 21; Poirion, D. "Ecriture et Ré-Écriture Au Moyen Âge." pp. 109–118 in *Littérature* 41 (1941); Cerquiglini, B. *Éloge de la variante*. Paris: Éd. du Seuil, 1989.

[70] Patten, McElligott, p. 14; Sentilles, R.M. "Toiling in the Archives of Cyberspace." pp. 136–156 in *Archive Stories: Facts, Fictions, and the Writing of History*, edited by A. Burton. Durham: Duke University Press, 2005; Eggert, p. 67; Jeanneney, p. 68.

[71] Ong, p. 78.

Ong alludes here to the idea that to refute a printed book, one needs to publish and print another book. Otherwise, it will not be on the same footing. To make counter arguments in writing, in the margin of one's personal copy, is rather pointless, as you know that there are hundreds, perhaps thousands of other copies without your marginal comments, and future copies will be based off of the files of the publisher, and not your commented-upon copy.

Meanwhile, digital documents allow users to write directly in the document itself. This constitutes different kind of **conversations**. In the manuscript world, we witness dialogues between the body of the text, the author, and the margin, the reader.[72] It is as though there is a person-to-person conversation going on.[73] This stipulates a style that is determined by building on earlier texts. It is normal for a text in the manuscript world to consist largely of copies of other texts, with only a small percentage of alterations, additions, and deletions. In printed texts, we have monologues, with the author telling the reader exactly how it is, which perhaps partly explains why the analytical framework of the author's intentions was dominant for most of the modern era.[74] Newspapers are a particularly good example of this, often filling the entire space of the page with walls of justified text. In case there is not enough text, newspapers rather insert 'fillers' than leave the space blank.[75] Of course, this does not mean there is not a trace of dialogue in the print world. There is,[76] just as much as the author can long be dead while readers still come up with new responses to his or her work.[77] But the motor that drives the print world does not fuel on such readers, only on genius authors.[78] Tim Ingold sees this as a result of the mechanization of text production. Likening sign language to other manual gestures, he says that "so long as the movement of the hand leaves an immediate trace on the page, there is no great difference between looking at signed words and looking at written ones."[79] Manuscripts, according to him, are inhabited with living voices, whereas "the voices of the past are eliminated from the printed text."[80] Indeed, it is a common complaint that the printed text 'do not talk

[72] Ong, p. 130.

[73] Pedersen, p. 35; Chia and De Weerdt, "Introduction," p. 18; Lit, L.W.C. van. "Islam Felsefesi ve Bilginin Dolayımi: El Yazmaları Üzerinden Yüz Yüze Sohbet." pp. 78–81 in *Sabah Ülkesi* 52 (2017).

[74] Finkelstein, p. 80.

[75] Ong, p. 130.

[76] Finkelstein, p. 108.

[77] Barthes, R. "The Death of the Author." *Aspen* 5–6 (1967).

[78] Wogan-Browne et al., p. 6; Ingold, T. *Lines: A Brief History*. London: Routledge, 2007, p. 24.

[79] Ingold, p. 28. Cf. pp. 93, 127, 139.

[80] Ingold, p. 24.

back'.[81] It is no surprise that the style of printed text is often determined by the adage 'say it once, say it well.'

In the digital world, I think there is essentially only soliloquy, a dialogue strictly with oneself.[82] In this sense, digital writing is the conclusion to what Walter Ong has perceived as a consequence of the separation of knower and known, namely that "writing makes possible increasingly articulate introspectivity."[83] This aspect of the digital world may be surprising, as it seems exactly the 'social' aspect that is often developed and highlighted.[84] But it is a consequence of the 'silofication' of the digital world, which separates each document or a bunch of documents from others without a clear context to place them in.[85] As far as there is interaction between people, this is primarily exploited to try something out and use the feedback to try something better since any act of writing has a trivial cost and writers can act as though they are their own publisher.[86] As such, it combines aspects of the manuscript and the print world: from the former, it takes the evolutive character of texts; from the latter, the one-directional decision making.

1.3.3 Product

There are four more notions to discuss on the level of the product itself. One is the end result of the **product**. A manuscript codex produces a process in flux. What I mean by that is that the entire production process can be witnessed, down to the mistakes that were crossed out and corrected, and often, it can generally be established what the order of that process was.[87] In chanceries, rather than throwing out drafts, they were crossed out.[88] In addition, a manuscript's result is open-ended; a reader is invited to continue the process by adding text (or subtracting by crossing out). Tim Ingold provides us a good image of how a manuscript works: "the surface of the page [is] like a country in which one finds one's way about, following the letters and words as the traveler follows footsteps or waymarkers in the terrain."[89]

[81] E.g. Blatty, W.P. *The Exorcist: A Novel*. New York: HarperCollins, 2011 [1971], pp. 36–37.

[82] Kathryn Sutherland hints at it. Sutherland, "Being Critical," p. 24.

[83] Ong, p. 103. Cf. McLuhan, p. 32.

[84] Merkoski, p. xii; Mandell, p. 155.

[85] McGann, p. 30. Weiss, p. 28.

[86] Cf. Müller; Finkelstein, p. 120.

[87] Johnston and Van Dussen, pp. 4–6; Dagenais, p. 17. Dagenais also says that "the handwritten text as product resembles the mechanically reproduced book; the process of its creation mimics the unique, occasional nature of oral tradition and oral performance." It is an interesting observation but I cannot do anything with it in my own analysis here.

[88] Vismann, p. 26.

[89] Ingold, pp. 24–26.

A printed publication, by contrast, presents the reader not with a process, but with a result.[90] Exactly how the text visible on paper came to be is entirely unclear from a mere examination of the page, which is why 'the' publication year is hailed as an important metric.[91] Ingold likens print publications to cartographic maps: "Had it not been for the journeys of travelers, and the knowledge they brought back, it could not have been made. The map itself, however, bears no testimony to these journeys."[92] Further, this is a fixed end result as it cannot be changed. Anything added to it will stand out as handwriting and, therefore, not part of the print world. If it does come out in the print world, it will be in a separate publication, remaining largely harmless to the original text.[93] This gives it a sense of veracity, which can be deceiving.[94] Digital documents also present the user with an end result, with no immediate way of knowing how the elements of the document came to look like they do. However, digital documents are in flux, as any reader can alter them without leaving any trace of such alteration. A next user would simply assume that the altered version is the actual end result of that digital document.

This difference in the end result tells us something about the different perceptions of the **fixity** of the product. An open-ended manuscript is known to shift and, therefore, not expected to be fixed. It is good to point out that this expected shifting can take different forms across different cultural zones. Notably, in the Islamic world, there is a certain reverence for authorial intent, and rarely do people take the liberty to overtly change the wording of a previous text.[95] In medieval Europe, however, such arbitrary interventions were commonly accepted.[96] If two manuscript copies of the same text show a difference, this ought to be considered a variant, not an error, as that is what makes them manuscripts and not a printed book.[97] When scholars want to transfer a text from the manuscript world to the print world, they have, all too often, defined their approach using the print world's terms, which proposes an illusion of being fixed, as though set in stone—typographic fixity, as Elizabeth Eisenstein has

[90] Ong, p. 129.

[91] Patten, McElligott, p. 12.

[92] Ingold, p. 24.

[93] Mandell, pp. 155, 174.

[94] This counts too for the Far East, see Chia and De Weerdt, "Introduction," p. 12.

[95] Lit, L.W.C. van. "Commentary and Commentary Tradition: The Basic Terms for Understanding Islamic Intellectual History." pp. 3–26 in *MIDEO* 32 (2017).

[96] Nichols, S.G. "What Is a Manuscript Culture? Technologies of the Manuscript Matrix." pp. 34–59 in *The Medieval Manuscript Book: Cultural Approaches*, edited by M. Van Dussen and M. Johnston, Cambridge: Cambridge University Press, 2015, p. 35.

[97] Dagenais, p. 18.

called it.[98] This is the proposal of the Lachmannian critical edition.[99] The use of such an edition, neglecting the manuscripts it emanated from, is seemingly taken for granted.[100] This counts as much for the 20th century as it does for the 10th, as long as it is a print world in which people operate. Thus, the Chinese ruler Mingzong demanded that even manuscript copies should be based on the printed versions of those texts, lest there be 'interpolation.'[101] The reality is, however, that even for born-print texts, differences between editions can exist, and even within one edition, copies can show variance due to the changes made during the process of printing.[102] Walter Ong notes the difference between the manuscript and print world: "Writing moves words from the sound world to a world of visual space, but print locks words into position in this space."[103] He notes a similar difference in the *mise-en-page*: "Chirographic control of space tends to be ornamental, ornate, as in calligraphy. Typographic control typically impresses more by its tidiness and inevitability."[104] Digital documents have an even bigger illusion of fixity since their existence relies on zeros and ones, and it therefore is, or it is not.[105] But what makes this an illusion is that the string of zeros and ones can at any time be changed. This is sometimes proposed as a desirable feature,[106] without taking into account the uncertainty this fosters. For example, documents can be silently withdrawn, and users can be banned without the possibility to appeal.[107] More importantly, since the supporting technology—both hardware and software—undergo a seemingly inexorable

[98] Eisenstein, E. *The Printing Press as an Agent of Change: Communications and Cultural Transformations in Early-Modern Europe*. 2 vols. Cambridge: Cambridge University Press, 1979, vol. 1, p. 116.

[99] Kleinlogel, A. "Variants and Invariants: The Logics of Manuscript Tradition." pp. 259–268 in *Theoretical Approaches to the Transmission and Edition of Oriental Manuscripts*, edited by J. Pfeiffer and M. Kropp. Beirut: Ergon Verlag, 2007.

[100] Dagenais, pp. 112, 114; Sutherland, "Being Critical," p. 25; Nichols, S.G. "What Is a Manuscript Culture," pp. 34–35.

[101] Chia and De Weerdt, "Introduction," p. 24.

[102] Lerer, S. "Bibliographical Theory and the Textuality of the Codex: Toward a History of the Premodern Book." pp. 17–33 in *The Medieval Manuscript Book: Cultural Approaches*, edited by M. Van Dussen and M. Johnston, Cambridge: Cambridge University Press, 2015, p. 18; Finkelstein, p. 20. Just consider the jungle of differences in the printed work *Vilette* by Charlotte Brontë, described in Sutherland, "Being Critical," p. 16.

[103] Ong, p. 119.

[104] Ong, p. 120.

[105] Baudrillard, p. 117.

[106] Eggert, p. 64.

[107] Jeanneney, p. 48.

change, the sustainability of digital documents in, say, one or two generations, is an acute problem which has been offered very little attention.[108]

2 Case Study 1: *ABC for Book Collectors* versus *A Dictionary of English Manuscript Terminology*

To better understand some of the differences between the manuscript world and the print world, and to see the previously discussed theoretical consequences played out in reality, let us compare two books that are each other's counterparts, one representing the manuscript world, the other the print world. For this purpose, we can make excellent use of *A Dictionary of English Manuscript Terminology 1450–2000* by Peter Beal and *ABC for Book Collectors* by John Carter and Nicolas Barker. Beal no less than opens his preface by saying: "This dictionary was originally inspired by John Carter's *ABC for Book Collectors* (first published in 1952). What he had done for books it seemed reasonable to do for manuscripts."[109] To compare the two books is, therefore, baked into the very plan of the books. As Beal's book was published in 2008, it seems good to somehow bridge the fifty-year gap between the two publications. To make the comparison as fair as possible, we shall avail ourselves of *ABC*'s 8th edition, which was prepared by Nicolas Barker in 2004 and corrected in 2006 by Raymond Williams.

Both books are a collection of entries, and 158 of them bear identical titles in both books. A majority of them were either exactly alike or did not reveal significant differences between the manuscript and print world. Twenty six of them, however, laid bare the theoretical differences previously discussed. I group these entries under three themes, concerning (1) the difficulties of transitioning from manuscript world to print world, (2) the craft of placing text on paper, especially regarding its decoration, and (3) the qualities desired by buyers.

2.1 *Manuscript Practices in a Print World*

Features typical of manuscripts are all of a sudden undesirable in printed works. ABBREVIATIONS, for example, were very common in manuscripts. Writing things by hand is labor intensive, and any shortcut is welcome.

[108] Jeanneney, pp. 63–64; McGann, p. 27. He notes that exactly our classic fields of "philology and textual criticism can help us gain that knowledge."

[109] Beal, P. *A Dictionary of English Manuscript Terminology 1450–2000*. Oxford: Oxford University Press, 2008, p. viii.

Abbreviations provide such a shortcut, all the more because a hand governing a pen offers a convenient and flexible tool to write, for example, "&" instead of "et." This flexibility is lost in the print world, when the typesetter has a limited number of type sorts to set any text. If anywhere the personal touch of a copyist is lost against the uniformity of typesetting, it is with abbreviations. Thus, a manuscript may read "oꝑa" with a dash through the bar of the *p*, by which is meant the Latin word *opera* (works). In printed works, this word is often typeset simply as "opera" or, in the case of some critical editions, as "op[er]a". As may be clear from seeing "oꝑa" restored to its former glory, in the digital world we are better able to encapsulate such glyphs. The important difference between "oꝑa" and "opera" is that in the latter case, spelling is dictated by the producers of the publication.

A similar notion can be witnessed in the production of ALMANACS. In the manuscript world, Beal points out that almanacs often had "blank pages or spaces for owners' personal entries to be made by hand."[110] There is, then, a notion of generosity and inclusion of the reader. For printed almanacs, in contrast, we read that they were "protected by jealously guarded patents,"[111] that is, they are marked by exclusion. A similar tension is witnessed in the entries on SAMMELBAND. Beal, writing about manuscripts, simply remarks that one volume containing different works is a known phenomenon. Carter, writing about printed works, notes the undesirability of this, and the frequent destruction of the binding and separation of the works; whereas texts live inclusively in the manuscript world, they generate exclusivity in the print world.

In this regard, the entries for ANONYMOUS authors are relevant too. Beal points out that "a huge number of literary works" are without a known author.[112] He goes on to describe how the anonymity of the author was of no problem in the manuscript world. Carter, meanwhile, mentions that "a book by an unidentified author is harder to sell."[113] What was a neutral phenomenon in the manuscript world, becomes contentious in the print world. This is, I would argue, because of the shift of the focus from the readers to the medium.

We can witness a shift of power from the copyist-reader to the publisher-medium. Whereas manuscripts have a COLOPHON at the end, describing mostly the context of the commissioning reader, printed works instead describe the context of the publishing author. Moreover, the print world phased

[110] Beal, p. 14.

[111] Carter, J., and N. Barker. *ABC for Book Collectors*. 8th ed. London: Oak Knoll Press, 2004 [Or. 1952], p. 23.

[112] Beal, p. 17.

[113] Carter, p. 25.

out the colophon at the end and replaced it with a TITLE PAGE at the beginning. Both the change in contents and the change in place show an affirmation of power by the publisher and related companies. One such change of contents is the ritualization of the date of issue. In the manuscript world, texts simply emanated from the author by the demand of readers for a copy, which made it important to mention the date of copy in the colophon. In the print world, the notion of PUBLICATION became all-important: "the offering of the book, for sale, to the public."[114] Thus, the date of publication is included on the title page. The very existence of a text is, furthermore, guaranteed much differently. For manuscripts, it is the general agreement among a body of readers that a text is worthy to be saved and copied. For printed works, it is the publishing industry who makes this decision. It is not without reason that even though the term IMPRIMATUR can be found in both Beal's and Carter's book, it nevertheless only applies to printed books, as a synthetic replacement for the consensus-process of the manuscript world.

It is, finally, striking that some terms become mere homonyms. FINGERPRINT, for example, refers in the manuscript world simply to the imprint of a human finger, left behind in manuscripts because of spilled ink. In the print world, it relates to a process to distinguish differences in typesetting by "comparison of several letters in adjacent lines."[115] Similarly, a LETTER in the manuscript world is a private message written by one person to another; whereas, in the print world it refers to the pieces of metal that can be typeset. In both cases, the meaning has shifted away from an individuating aspect of the copyist to an identifying aspect of the publisher.

2.2 *Fear of Voiding or Fear of the Void*

Horror vacui, or fear of the void, is a phenomenon well-known to any manuscript culture. As the great medieval litterateur al-Jāḥiz says, "a black space is better for it than a white one."[116] Marginal notes are the bread and butter of the manuscript world. All of a sudden, this becomes different in the print world, where there is a fear of voiding: it is exactly the 'staining' of the pages, including the white space, that is avoided. We see this reflected in several entries from Beal's and Carter's books. ERRATA provides a good starting point. Whereas manuscripts would be checked for accuracy and corrections would be made in the margin directly at the offending place, corrections to printed works would

[114] Carter, p. 180.

[115] Carter, p. 103.

[116] As translated by F. Rosenthal in *The Technique and Approach of Muslim Scholarship*. Rome: Pontificium Institutum Biblicum, 1947, p. 6.

be made on a separate piece of paper and either bound with the book or simply inserted as a loose sheet. The offending page remains untouched. The term COLLATION, therefore, means something entirely different in the two worlds. Beal describes it as the process of checking the text of one copy against another, whereas Carter describes it as the process of checking the division of pages into quires of one copy against another. For manuscripts, then, the emphasis is on how the copyist rendered the text, whereas, for printed works, the emphasis is on how the publisher and printer rendered the volume. Furthermore, while the collation for manuscripts is, in a certain sense, a neutral process, leaving open the correctness of a reading, the collation of printed works is to judge the completeness of a copy. This completeness, of course, is about how the book was produced, not whether the full text is available in the volume. This is further reflected in the difference in using the term VARIANT. For manuscripts, a variant means "a reading in one or more manuscript or printed publications that is different from the one in the particular text under scrutiny."[117] For prints, a variant is "to describe a copy or copies of an edition or impression exhibiting some variation, whether of text, title page, illustrations, paper or binding, from another copy or copies of the same edition or impression."[118] In other words, while the manuscript world uses the word variant to indicate a difference in contents, the print world uses it to indicate a difference in medium.

HISTORIATED is another good example of the difference in fear. Historiated means that the initial capital is decorated with, properly speaking, a scenery but any decoration could work. In the early stage of the print world, such decorations were expected to be done, and the typesetter would only print a GUIDE LETTER, a small letter to indicate the letter which the decorator should draw and decorate. It seems that quickly these letters were left standing on their own and no decoration was applied; whereas manuscripts would be labored over for hours to decorate, printed works would be left untouched.

The term DECORATION itself provides for a striking comparison. Whereas Beal speaks of the decoration of the text on the page, Carter defines decoration in terms of the embellishment of the volume, the book as a whole, especially the covers and the binding. The 'void' of the white space on the pages is left alone, perhaps out of a fear of voiding the pristine quality of the copy of the book. No other aspect of manuscript and print production shows this difference more starkly than ILLUMINATION and GILDING. Whereas for manuscripts this means the use of gold leaf either to write out text or to embellish it, for print works it means the decoration of a book's edges when closed flat. It is

[117] Beal, p. 428.

[118] Carter, p. 227.

as though the page itself is the domain of the publishing industry, which is not to be messed with, and so the gilding literally falls off the pages!

2.3 *Moldy Manuscript or Pristine Print*

Lastly, in their description of manuscripts and printed works, Beal and Carter demonstrate that a different aesthetic is at work. This speaks firstly from Beal's stereotypes of the greatest lovers of manuscripts and printed works, respectively. He uses John Earle's 1620s description of a manuscript collector "as a wrinkled old man who 'loves all things … the better for being mouldie and wormeaten.'"[119] A collector of printed works, meanwhile, is described as someone "who is devoted to acquiring or assembling a collection of books or manuscripts, generally as artifacts in their finished form."[120] The contrast of the two is not so clear at first, but the crux lies in the term 'finished form.' Whereas manuscript lovers do not mind a good WORMHOLE, Beal informs us that such is "universally reviled by librarians and book collectors."[121] Finished form, then, is a synonym here for pristine. Carter literally glosses 'finished product' with "the product in its first, pristine form."[122] 'Pristine' is a good word to describe the desired aesthetic of a book, for it hints at both an unspoiled state, clean and crisp, as well as an original, untampered state. The two aspects of pristine sometimes conflict with one another. "The greatly increased respect for original state among collectors," says Carter, "has tended to reduce to the minimum the amount of tampering with even battered copies."[123] Clearly, then, whereas for manuscripts the wear and tear caused by readers is a desirable quality, for printed books, it is the original state of publication that is sought out.

Other forms of tampering or wear and tear are similarly different. Under the entry INSCRIPTIONS, Beal describes in relatively favorable terms how manuscripts often have additional handwritten comments. Carter, on the other hand, speaks of them in terms of "something of a defacement."[124] Those copies from a library (that is, EX-LIBRARY) that show such traces are valued by Beal as "obviously valuable evidence of provenance,"[125] whereas Carter writes that such "traces are regarded with lively disfavor by most experienced collectors and with contempt by the fastidious."[126] The entire BINDING of

[119] Beal, p. 19.
[120] Beal, p. 80.
[121] Beal, p. 441.
[122] Carter, p. 181.
[123] Carter, p. 190.
[124] Carter, p. 131.
[125] Beal, p. 148.
[126] Carter, p. 96.

a volume is itself subject of aesthetic difference. Manuscripts, Beal writes, "are likely to be in bindings supplied by subsequent owners, reflecting their individual tastes or circumstances."[127] I did not find a similar comment about bindings of printed works in Carter's book, but the presence of an entry called 'Deckle-Fetishism', "The over-zealous, undiscriminating passion for uncut edges in books which were intended to have their edges cut"[128] should say enough.

The difference in aesthetic is, lastly, rather splendidly demonstrated in the differences between the entries on Facsimile. Beal has a rather tame and neutral description of what a facsimile is: "an exact copy […] imitating in every detail the original physical artifact."[129] Given what we know of the manuscript world, though, I think we may surmise that a facsimile is something that is quite desirable, a dream of any copyist. We may think of the many facsimiles (in the form of printed and bound photos) in use by scholars of famous manuscripts, allowing us to examine the manuscript 'as it is' without being on site. In the print world, according to Carter, "it figures frequently in the nightmares of collectors, [and] causes booksellers more trouble than almost any other factor in their business."[130] The reason for this is quite clear. "An exact copy is a menacing thing to those who pursue originals."[131] I think demand drives supply here. The apparently real market of forgery of rare books shows how desperate the craving for pristine copies is. Seeing that a book is mechanically produced, what would be the disadvantage of having a mechanical reproduction? The text contained in it is presumably the same so the reading experience will be about equal. The problem becomes real if it is the medium that is central to the print world, since it is exactly the medium that is being tampered with in the case of a facsimile. Dismissing a facsimile is neither about the authorial intent nor the readers' response, but the medium itself and the authority it derives from its publisher. A facsimile steps into that space of the medium but without authorization and, as such, it chips away at the very epistemology of the print world. Book collectors, it seems to me, intuitively understand this; hence, their outrage.

[127] Beal, p. 39.
[128] Carter, p. 79.
[129] Beal, p. 150.
[130] Carter, p. 98.
[131] Carter, p. 99.

3 Case Study 2: *A World Without Whom* versus *Do I Make Myself Clear?*

Let us make a concrete comparison between the worlds of print and digital. Again, I merely want to point out some notable differences, mostly having to do with their assumed epistemology since that is what largely constitutes their world-making potential. I found an excellent couple in Emmy Favilla's *A World Without "Whom"* and Harold Evans's *Do I Make Myself Clear?*[132] One is a young copy editor at BuzzFeed, the other a veteran editor of *The Times* and *Sunday Times*, currently at Reuters. The books are highly comparable, as both provide a writing guide, and both decided to write a guide because so much has changed with the rise of the rivalry between digital writing and printed writing. One defends a way of writing associated with print; the other merrily advances the case for a new way of writing—the way of the internet. They were both published in 2017, drawing, now and then, from the same source material. Whereas the book defending print-writing uses ample examples from the internet, uses the word "listicle" and even includes in the body of the text a URL to a YouTube video, the book defending digital writing all too often shows its reverence for Merriam-Webster Dictionary and the Associated Press Stylebook. What, finally, makes these books clearly a good combination for comparison is that it would be hard to decide which one is which, merely based on the titles. It is *A World Without "Whom"* which is defending a new way of writing, native to the digital world, which welcomes the disappearance of the word 'whom.' But one could easily imagine it to be the title of the book defending a writing style native to print, bemoaning its disappearance.

3.1 *Word of the Year: 2015 or 2016?*

Favilla and Evans have different reasons for writing. In as much as their reasons come together, they are each other's opposite. This can be epitomized by their shared appreciation for the 'Word of the Year' as chosen by the Oxford English Dictionary. Favilla feels vindicated by 2015's word, which was chosen to be, controversially, the "Face with Tears of Joy" emoji. As Favilla puts it: "I mean, what a time to be alive, seriously."[133] At the same time, she argues herself that an emoji is not a word, and the emojis featured on the cover of her book are purely decorative. This can be understood when we read on the

[132] I was pointed in this direction by Tom Rachman's review of them in "Writers gonna write" pp. 8–9 in *The Times Literary Supplement*, no. 5990, January 19 2018.

[133] Favilla, E.J. *A World Without "Whom."* London: Bloomsbury, 2017, p. 260.

first page that "this book is about feelings."[134] Emojis do not need a precise definition; as long as they convey a certain feeling or aesthetic, they are warmly welcomed by Favilla. In other words, she is aware that the digital way in which she has been using language supports the idea that texts are supposed to be subjective.

Evans, however, attaches greater value to 2016's Word of the Year, which was 'post-truth.' To him, "we are certainly in the vortex of what's come to be called the post-truth society."[135] Evans parses this notion to a figure of speech, that of 'fog.' "Fog everywhere. Fog online and in print,"[136] he opens the book. And he finishes the book saying that "The fog that envelops English is not just a question of good taste, style, and aesthetics. It is a moral issue."[137] This is because, in Evans's words, "Words have consequences,"[138] and he contends that "the oppressive opaqueness of the way much of English is written is one cause for a retreat from reason to assertion."[139] The enabler of this opaqueness he finds in "digital social media," which is "an unwitting agent for the mass dissemination of rumor and semi-truth now known as Fake News."[140] He concludes that "the consequences of all the propaganda are real,"[141] and gives the 2016 US Elections as an example. In other words, Evans is uneasy with a perceived erosion of language's use as a conveyer of truth. In the print way that he has used language for his entire life, the text is supposed to be objective.

The subjective/objective dichotomy is reflected in how I argued that the digital world has the sender (authors) at its core and the print world has the medium (pages). This constitutes a different power dynamic. Thus, Favilla does note OED's 2016 Word of the Year (post-truth) but dismisses it as irrelevant. When it comes to writing, she says that "you can start your sentences however you please, because your stylistic preferences make you you. Also? They're words, not weapons."[142] This is in stark contrast to Evans, who quotes Christine Kenneally for making the point that as a writer, "you are a god in language. You can create. Destroy."[143] For Favilla, a single digital text is innocent, whereas Evans attaches great weight to it. Favilla says: "Select the phrasing

[134] Favilla, p. 1.
[135] Evans, H. *Do I Make Myself Clear?* London: Little, Brown, 2017, p. 194.
[136] Evans, p. 4.
[137] Evans, p. 347.
[138] Evans, p. 3.
[139] Evans, p. 16.
[140] Evans, p. 302.
[141] Evans, p. 306.
[142] Favilla, p. 136.
[143] Evans, p. 346.

that gets the intended meaning across in the briefest way. You've got the power here, and I'm confident you'll use it wisely."[144] Such a statement betrays a mentality in which one text cannot possibly make a big difference. As for the power that an author does have, they have it as a birthright in the digital epistemology. This shines through in Favilla's book when she says that "artistic license is especially constructive when the internet is the medium."[145] This is because digital writing is "often more personal and more plan-languagey."[146] If the meaning of words is corrupted along the way then "that's neither sad nor cause for outrage—it's simply reflective of how the word has devolved."[147] Most telling is the following passage from Favilla's book:[148]

> Since we live in an era where we can literally TALK TO THE DICTIONARY, we decided to go straight up to the source and ask what the hell was up.

The passage is followed by an image of a tweet directed at Merriam-Webster, questioning the veracity of a dictionary entry. The tone in the sentence above is purposefully rude, as a joke to contrast the seriousness of doubting the judgment of a respectable publication such as Merriam-Webster Dictionary with the easiness with which one can, nowadays, publicly make known such a doubt. In other words, Merriam-Webster might rule over what is printed on paper, thereby consolidating power through the medium, but in a digital world such power leaks away, since anybody can write back and call into question the very definition of words.

Meanwhile, print epistemology is defended by Evans. It is as though he directly responds to Favilla's rude statement about calling into question dictionaries when he says: "The gatekeepers who scrutinized all entrants to the citadel of print have now been outflanked by 0s and 1s in the millions."[149] What can be said in a manuscript world is decided by readers, owing to their decision regarding the texts to be copied. What can be said in a print world is decided by 'gatekeepers'—the publishers and their auxiliaries, through their control over the 'citadel'—the medium of the page. What can be said in a digital world, however, is decided by writers. Each may only be insignificant, but add the millions together, and discourse will gradually move in a certain direction. What we argued for in the theoretical part of this chapter we can now see confirmed

[144] Favilla, p. 155.
[145] Favilla, p. 1.
[146] Favilla, p. 17.
[147] Favilla, p. 227.
[148] Favilla, p. 84.
[149] Evans, p. 197.

in the wild, by Favilla, who affirms that "today everyone is a writer—a bad, unedited, unapologetic writer."[150] For those who grew up in a digital world, this is a rather obvious statement, barely worth making. But for a person like Evans, born and raised in the print world, it is a fact he cannot even muster the patience for. Addressing his readers, he says, with misplaced irony: "If you are more into explaining your inner self to the waiting world than in conveying information, stop here and brood on to greatness."[151]

3.2 *Thinking to Type or Typing to Think? Typewriter versus Text Messages*

The theoretical differences I noted in terms of process and the end product can be seen in the books by Favilla and Evans. Both present one guiding principle that tells a lot about their assumed epistemologies. For Favilla, it is "break any of these rules sooner than say anything outright barbarous."[152] This is notably a destructive, negative principle, one that breaks rather than builds up. It further does so by passing judgment on the thing broken, namely, it is something barbarous. No caution is advised, and as such, the principle seems egotistic. This becomes all the more clear when we compare how Evans phrases the same advice (though not his guiding principle): "we should respect grammatical rules that make for clarity, but never be scared to reject rules that seem not to."[153] Here, the onus is on the writer to overcome hesitance, fear even, in allowing themselves a lapse of a rule, always accompanied with respect for those rules. This, inter alia, translates into diametrically opposing sentiments on the classic writing guide *The Elements of Style* by William Strunk and E.B. White; for Favilla, it is her favorite representative of the print world to bash, for Evans it is his gold standard for which to aim.

Meanwhile, the guiding principle of Evans is: "Pity the reader."[154] The principle seems mostly to slow authors down and make them think before they write. It is, notably, a top-down statement, one that can easily turn into a paternalistic attitude. The closest equivalent in Favilla's book is when she writes that "it's often crucial to be mindful of respectful and inclusive terminology,"[155] which has a more horizontal, peer-to-peer tone to it.

The writing process that those guiding principles encourage is sharply different. Favilla boldly admits that "the relative ephemerality of modern written

[150] Favilla, p. 16.
[151] Evans, pp. 20–21.
[152] Favilla, p. 20.
[153] Evans, p. 21.
[154] Evans, p. 19.
[155] Favilla, p. 22.

communication means that there's simply less thought that goes into the words we toss onto a screen."[156] She celebrates this, saying that "we are no longer slaves [...] in front of a rickety old typewriter."[157] Her choice for a typewriter to symbolize anything that would slow us down is interesting, as it is the same symbol that Evans upholds as the ideal writing process. He fondly invokes an image of E.B. White, author of *The Elements of Style*, writing: "the melodies White made at his typewriter—hesitant bursts of *clack-click-ring*, with long silences in between and then brooding silences at lunch, worrying about the words he left unfinished."[158] Clearly, for Evans, writing should not be ephemeral but an exercise in patience. A typewriter offers patience by making every keystroke a considered decision.

Favilla and Evans consider different aims for writing texts in a digital world. Evans, considering the consequences of a post-truth society, provides a solution typical of print. He argues that "the maelstrom of mendacity makes it all the more imperative that truth be clearly expressed,"[159] to which he elsewhere adds that "cyberspace is indulgent, but attention spans are shorter. We appreciate conciseness."[160] Evans, then, aims for short and correct writing, which, indeed, can only be achieved if one meditates on their writing. Favilla, meanwhile, aims for an "acute connection with readers that drives them to engage with and share your content."[161] Truth and brevity can take a backseat because writing "has to scream over the crowd to get the views, the likes, the shares."[162] Avoiding untruth only matters for Favilla insofar as it would hurt the trust readers have in you. Even then, it is not about stating something untrue, but it "can lead to a lack of respect for stories you produce *in the long run*."[163] Favilla's head is already with the next piece; the engagement you generate now is as an investment to build on and grow in the future.

This, finally, results in two distinctly different envisioned products. Favilla touches on this when she says that "any form of social media allows us to indicate the end of a sentence by pressing 'send.' Unlike analog generations of yore, we often send thoughts piecemeal, rather than as a complete package."[164] In the first instance, this statement relates to personal digital communication,

[156] Favilla, p. 154.
[157] Favilla, p. 153.
[158] Evans, p. 320.
[159] Evans, p. 15.
[160] Evans, p. 226.
[161] Favilla, p. 20.
[162] Favilla, p. 226.
[163] Favilla, p. 81, emphasis added. Cf. p. 36.
[164] Favilla, pp. 244–245.

such as text messages. However, I would argue, it also applies to public statements such as tweets, blog posts, and online articles. Favilla is most concerned about the credibility of authors in terms of the likeliness of readers to read the *next* story, and all that is required *right now* is an 'acute connection with readers.' Thus, it is more important to get somebody's attention with an unfinished thought as soon as possible than to sit on it and work on refining and completing that thought. There is a similarity here with photography and art. One photograph, we know, does not constitute art, it being merely a mechanical reproduction with little room for artistic intent. But a series of photographs can be art. Similarly, one born-digital text is hardly a finished product, but a series of digital texts can. If you do not get it right the first time or failed to capture an audience, simply say it differently. Text itself is disposable, as the author will continue to convey the message.

The desired product of Evans is neatly captured in a cartoon which he includes in his book. It depicts a casual reader holding a freshly written sheet of paper, standing next to what looks like Shakespeare, who is steaming with anger, a caption reading "Good, but not immortal."[165] Evans included it obviously as an ironic comment, but the irony can only work if we actually believe that writing should strive for immortality. Indeed, Evans speaks of "great writing"[166] as our aim and is constantly concerned with bringing down the word count. In short, he is of the school 'say it once, say it well.' Text, thereby, becomes invaluable, carrying the meaning of the author's thoughts independently.

4 Case Study 3: The Written, Printed, and Digital Koran

Some literature has co-existed with people since time immemorial. The *Alexander Romance*, for example, or *Kalila and Dimna*, have been alive for more than two millennia among a great number of cultures and languages. In fact, these stories are very fluid, breaking down in all kinds of versions and subversions, depending on the time, place, language, and undoubtedly other factors. As a consequence, these stories find expression in all the different worlds; manuscript, print, and digital. Other writings need to be retold, too, but they cannot be fluid since they are considered sacred. Examples are the Daoist *Tao Te Ching*, the Christian *New Testament*, and the Islamic *Koran*. Those with knowledge and vested interests in these texts consider themselves as custodians, taking great care in finding the right balance between assuring the

[165] Evans, p. 176.

[166] Evans, p. 91.

continued existence of these texts and avoiding aberrations. Observing how such a text moves from one world to another is, therefore, an interesting test case for witnessing the particularities of each world. In this case study, I focus on the Koran, as this holy scripture falls within my field of expertise.

4.1 *From Manuscript to Print*

Even though it is said that Muhammad received revelations orally, they were, from the very beginning, written down and compiled into a book known as the Koran. In fact, the Koran has multiple self-references as a 'book', which makes its identity as a written text all the more easy. One could go even further and argue that the written Koran was what made Islam a civilization of manuscripts more than any other.[167] The world-making potential of Islam is, I would argue, intimately tied up with the manuscript culture.[168] It is no wonder, then, that print technology was very slow to be accepted and adopted by Muslims. The first printing of the Koran happened, therefore, not by Muslim hands but Christians in Europe, in Basel and Venice in the 16th century.[169] Mention is made of an edict (*firmān*), supposedly issued by Beyazid II in 1485, which would ban any use of the printing press by Muslims.[170] However, I have been unable to find actual evidence for this order.

The first concerted effort in bringing print into the Islamic world, around 1729, was done by Ibrāhīm Müteferrika, who was of Hungarian-Christian decent but lived most of his life as an Ottoman Muslim. He assured permission from the sultan and the *shaykh al-islām*, the highest religious authority, together with confirmation of sixteen religious authorities, all active or former judges (sing. *qāḍī*). Additionally, he wrote an essay entitled *Wasīlat al-tibāʿa*

[167] Ahmed, S. *What Is Islam? The Importance of Being Islamic*. Princeton: Princeton University Press, 2015.

[168] Cf. Van Lit, "Commentary and Commentary Tradition.".

[169] Bobzin, H. "Von Venedig Nach Kairo: Zur Geschichte Arabischer Korandrucke." pp. 151–76 in *Sprachen Des Nahen Ostens Und Die Druckrevolution. Eine Interkulturelle Begegnung*, edited by G. Roper, D. Glass, and E. Hanebütt-Benz. Westhofen: WVA Verlag Skulima, 2002. Arjan van Dijk reconstructs, with some speculation, that the Venice print was meant for export to the Islamic world. After the Ottomans found mistakes in it, it was completely destroyed and the book seller's right hand was chopped off, cf. Dijk, A. van. "Early Printed Qur'ans: The dissemination of the Qur'an in the West." pp. 136–143 in *Journal of Qur'anic Studies* 7, no. 2 (2005).

[170] Larsson, G. *Muslims and the New Media: Historical and contemporary debates*. Farnham: Ashgate, 2011, p. 33; Oman, G. "Maṭbaʿa", *EI*², vol. VI, p. 795a; Leemhuis, F. "From palm leaves to the Internet." pp. 145–62 in *The Cambridge Companion to the Qur'ān*, edited by J.D. McAuliffe. Cambridge: Cambridge University Press, 2007, p. 152; Abdulrazak, F.A. "The Kingdom of the Book: The History of Printing as an Agency of Change in Morocco between 1865 and 1912." PhD dissertation Boston University, 1990, p. 76.

about the benefits of printing. We know all of this because Müteferrika inserted these documents at the beginning of the first work he printed.[171] The *firmān* and the essay have been translated by Christopher Murphy.[172] The *fatwa* was translated in English by Skovgaarden-Petersen.[173] The essay is also available in a modern edition together with a translation and analysis in Persian.[174] The decree from the sultan, the opinion from the judge, and the essay from the printer all emphasize that what is asked for is the printing of non-religious texts. Clearly, printing a sacred text was seen as a bridge too far.

It will be beneficial to draw from Müteferrika's essay to understand the ramifications of printing the Koran.[175] He outlines ten benefits, many of which betray the fundamental differences between the manuscript and print world. Benefit two and three talk about the pristine nature of print. A printed book, according to him, is "stable and enduring," and that if old texts are printed, they are "being restored and invigorated as if they had been recently authored."[176] He connects with these thoughts the suggestion that printed books are "safe from mistakes." So, we see that already at this earliest stage of the print world coming about in the Islamic world, print is seen as a pristine end product that definitively replaces moldy manuscripts.[177] Furthermore, the touchstone of veracity is no longer a long manuscript tradition of many readers selecting the best readings and copying it for future readers, but the publisher. As a fifth benefit, Müteferrika makes mention of page numbers, which he says facilitates immediate and accurate access to a passage, especially when a table of contents and an index is included in the back of the book. Texts until that time generally

[171] Jawharī. *Tarjama ṣiḥāḥ al-Jawharī*. Translated by Vānqulī, printed by Ibrāhīm Müteferrika. 2 vols., Istanbul: Dār al-ṭibaʿa, 1141h (1729), pp. i–xvi [pages not numbered]. These texts (except the consent notices of the judges) are in Ottoman Turkish. The relevant pages have been digitized at least twice: Budapest: National Széchényi Library, H 3252; Qatar: Qatar National Library, PJ6636.T8 J39 1729. The digitized version at McGill does not have these pages and a librarian confirmed that the copy itself does not have them: Montreal: McGill University Library, PJ6620 J382187 1729.

[172] Atiyeh, G.N., ed. *The Book in the Islamic World: The written word and communication in the Middle East*. Albany: SUNY Press, 1995, pp. 284–285 and pp. 286–292.

[173] Skovgaarden-Petersen, J. *Defining Islam for the Egyptian State: Muftis and fatwas of the Dar Al-ifta*. Leiden: Brill, 1997, p. 73. Also in Larsson, p. 32.

[174] Dhawqi, F. "Ibrāhīm Mutafarriqa, Risāle wasīle-ye al-ṭibāʿa wa-tarjama ān." pp. 234–282 in *Payām-i Bihāristān* 2, no. 4 (2016).

[175] A text closer to the actual printing of the Koran, by Muhammad Haqqi, is mostly a rehash of Müteferrika's arguments, see Abdulrazak, p. 89.

[176] Much more could be said about the impact of this new perception of texts. For now, let me do with a reference to Robinson, F. "Technology and Religious Change: Islam and the Impact of Print." pp. 229–51 in *Modern Asian Studies* 27, no. 1 (1993), especially p. 242.

[177] Cf. Nichols, "Introduction," p. 3.

had no page numbering and no index, only a rudimentary table of contents. Müteferrika, then, intuitively understood the immense centrality of the page number for the print world. Thereby, he hints at using a text completely differently from a manuscript to a printed book. Whereas one needs to be introduced to a text in a manuscript gradually and intimately, a printed book can be commanded by a reader much easier. As a seventh benefit, he notes that print publications can "become a foundation for the strength of the empire." In other words, Müteferrika understands how print can centralize power and can forcefully address specific classes of people. He also sees this power cast back onto the ruler—that is to say, that people will actually consent more easily to the government if they display the power of the printing press (benefit eight). He further points out that if the government does not allow Muslims to step into this position of power, others will, as European traders are already knocking at the door wanting to sell print publications (benefit nine). Benefits one, four, six, and ten tie into each other, putting forward the aspect of print that the archive does not attain a status of great power by quality (as manuscripts would, evidenced by the stance of Müteferrika's opponents) or accessibility (as digital documents would, evidenced by altafsir.com's *About* statement), but by sheer quantity. With this, additionally, Müteferrika emphasizes the educational potential of producing many cheap books.

A century later, in 1833, at the famous Bulaq Press in Cairo,[178] it was especially the educational argument that was used to initiate the printing of parts of the Koran. By then, print had proven itself in the Islamic world, or it at least had found a stable place alongside manuscripts. The time was ripe, it seemed, to take the next step and print religious writings. We have a partial record of the back-and-forth between the religious authorities and the printer regarding the project to print excerpts from the Koran. The clergy asked if any parts of the materials used for printing were made of dog skin.[179] Apparently, it is this materialistic aspect that interested the clergy, being sensitive to the different nature of making the word (and the name) of God come to be by means of a human being writing it by hand and by a machine stamping it. Since apparently no dog skin was involved, the printing went through. The clergy had already been worried about mistakes creeping in, and upon noticing mistakes, they asked for the books to be seized and its sale forbidden. They realized that

[178] Scholars have mentioned earlier prints, including from St. Petersburg, Teheran, Shiraz and Calcutta, but for our case study the Bulaq-episode will suffice. See Bobzin, pp. 166–167; Albin, M.W. "Printing of the Qur'ān." In *Encyclopædia of the Qur'ān*, edited by J.D. McAuliffe, 6 vols., Leiden: Brill, 2004, vol. 4, pp. 264b–276b.

[179] Ridwan, A. *Ta'rīkh maṭbaʿa būlāq*. Cairo: Bulaq, 1953, p. 279

mistakes in a print publication could have a significant effect on the overall perception of what the text ought to be, as all of a sudden there was not one copy with a mistake in it, but hundreds and, possibly, thousands of copies.

4.2 *From Print to Digital*

It was only in 1924 when a printed edition of the Koran was produced that would make a significant impact. I am talking about the Cairo edition, also known as the King Fuʾād edition, royal (*ʿāmiriyya*) edition, or Azhar edition. This printed version of the Koran was eventually accepted by virtually all denominations as a gold standard, from the Wahhabi Saudis to the Twelver Shiʾi Iranians.[180] Most notably, since the eighties, the King Fahd Holy Qurʾān Printing Complex churned out millions of copies each year of what they call the *muṣḥaf al-Madīna*, which is equivalent to this Cairo edition. The sheer quantity of this print production has made the medium—the Cairo edition as an abstract idea of a perfect text—the message. One example of this is the extraordinary rise of the Koran's talismanic use. Before, only one or few verses would be engraved, inscribed, or otherwise carried with, but now it is fairly common to see full miniature Korans hanging in cars, on keychains, or printed credit card-size or engraved on a medallion.[181] The *Alifī Qurʾān* is another such example, in which the text is typeset to begin every line (on every page) with the letter *alif.* Only in obscure sub-sub-sects of Islam is manual copying still obligatory.[182]

This one print edition stands in sharp contrast with the seven accepted compilations of the Koran in the manuscript world, which spawned seven additional approved versions, making fourteen readings in total. The variety of these readings is available in print, for sure,[183] but this seems to be lost on the print mind-set of most people. For, even though the Cairo edition does say at the end that it is only the text according to one reading, that of Ḥafṣ, transmitted as by ʿĀṣim, it "more or less eclipsed other readings" as Fred Leemhuis

[180] Indeed, even the Ibadis from Oman follow the Cairo edition. I make this claim anecdotally and realize a serious grand comparison is long overdue.

[181] Cf. Larsson, p. 173; Hirschkind, C. "Media and the Qurʾān." vol. 3, pp. 341b–349b in *Encyclopædia of the Qurʾān*, edited by J.D. McAuliffe, 6 vols., Leiden: Brill, 2004.

[182] Akkerman, O. "The Bohra Dark Archive and the Language of Secrecy: A Codicological Ethnography of the Royal ʿAlawī Bohra Library in Baroda." PhD dissertation Freie Universität. Berlin, 2015.

[183] See Mukhtar Umar, A., and A. Salim Mukarram (eds.). *Muʿjam al-qirāʾāt al-qurāniyya.* 8 vols. Kuwait: Dhāt al-salāsil, 1988.

puts it.[184] So much so that even scholars will refer to it as “the standard text,”[185] and some scholars find it problematic that the Koran in the digital world does not fully cohere to it.[186] Noticeably, websites hosting the Koran rarely mention what reading they use. For example, quran.com relies on the text provided by tanzil.net, which justifies its text by stating: “Tanzil Quran text is carefully derived from Medina Mushaf, which is currently the most authentic copy of the holy Quran (narration of Hafs).” Thus, even though there is a faint notion of the fourteen readings at the very end by referring to Hafs, the main belief propagated is that there is one authentic rendering of the text, the Cairo edition, to the exclusion of the rest. In this, we can see the quintessential spirit of the digital world to have a blinding seduction of usability. As altafsir.com explains its own purpose, it wants to be “instant, safe, user-friendly and easy.”

Similarly, whereas in the manuscript world there was a variety of names given to the different chapters (sing. *sūra*), and verse (*āya*) numbering was either absent or done in different ways, the Cairo edition solidified both.[187] The digital world brought this to its logical conclusion, namely, it *defines* the Koran through numbers. For example, the website quran.com uses a URL structure according to “quran.com/X/Y” with X as chapter number and Y as verse number.

4.3 *From Digital to Manuscript*

The actual entrance of the Koran into the digital world went by and large unnoticed. Contested issues had already been ironed out with the printing of the Koran and its release on audiovisual media.[188] Hence, access to the Koran had become thoroughly egalitarian, and its spread over the internet was not seen as a problem. Leemhuis reports a simple metric about the Koran in digital form, by giving the number of hits he got on Google when searching for several spellings of the word Koran.[189] Below, I contrast his results with mine, showing an explosive fifteen-fold growth.

[184] Leemhuis, p. 152.

[185] Puin, G.-R. “Vowel Letters and Ortho-Epic Writing in the Qur’ān.” pp. 147–90 in *New Perspectives on the Qur’ān: The Qur’ān in Its Historical Context 2*, edited by G.S. Reynolds. London: Routledge, 2011.

[186] Rippin, A. “The Qur’ān on the Internet: Implications and Future Possibilities.” pp. 113–26 in *Muslims and the New Information and Communication Technologies*, edited by T. Hoffmann and G. Larsson. New York: Springer, 2013.

[187] Most notably, Flügel’s edition shows differences, cf. Flügel, G. *Corani Textus Arabicus*. Leipzig: Sumtibus Ernesti Bredtii, 1869.

[188] Larsson, p. 184.

[189] Leemhuis, p. 158, fn. 13.

TABLE 1.4 Number of hits for keywords concerning the Koran

	2005	**2018**
Koran	6,440,000	38,700,000
Quran	3,890,000	108,000,000
Qur'an	1,400,000	39,900,000
Qur'ân	864,000	260,000
Bible	34,800,000	485,000,000
Koran/Bible Ratio	0.362	0.385

It should be noted that this is an overly simple metric, as there might be overlap in hits between "Qur'an" and "Quran," and there are other ways of spelling this word, and most notably we here only compare the English results of the word and not those in Arabic, Persian, Urdu, Turkish, Baha Indonesia, and all the other languages in which one would expect a major online presence of relevant materials for the Koran. Comparing to the Bible on absolute numbers is therefore not relevant, since the number of anglophone people who engage with the Bible is simply larger than those who engage with the Koran. However, we can take the term *Bible* as a control term. Considering that the ratio stays nearly the same (the Koran slightly winning terrain), it seems this metric does have saying power, and that there is truth in saying that the online presence of the Koran has grown fifteen-fold.

In the previous section, we saw how the digital Koran derives its existence from the printed Koran. However, interestingly enough, among the growth of the digital Koran is a noticeable looping back to its manuscript. The work of Thomas Milo is a prime example in this regard, in particular, two projects of his.

The first one is called *Qur'ān Concordance* and can be considered as a minimalist digital representation of the Koran. This product stemmed from Milo's observation that anything (even diacritics) beyond the actual lines of the words, the skeleton (*rasm*) or archigrapheme, was added later. Therefore, he created a digital encoding of just that skeleton, in order to get a more robust representation of what we can know for certain the Koran to have been at its earliest stage. He arrives at this skeleton by taking the Cairo edition and comparing it with the photos of Koran's early fragments. The observable dynamic between the digital and manuscript world is, then, notably different from the dynamic between the print and the digital world. Whereas when the print and the digital world meet, they together strive for a singular, unambiguous encoding of the Koran, when manuscript and the digital world meet, they open up the possibility for a multi-interpretable encoding of the Koran that can only reach a human-readable state by computation. Accuracy is a concern for both, but by comparing the encoding to manuscript evidence, accuracy becomes a test for the integrity of manuscripts, not an assumption about the unicity of the text.

Milo's second project is *Muṣḥaf Musqaṭ*, which digitally renders the Koran in a maximally detailed form.[190] This project relies on the 'Advanced Composition Engine' (ACE) developed by his company DecoType, which can typeset a script in a more flexible way than conventional font technology. An example of the advantage of *Muṣḥaf Musqaṭ* over the typesetting of the Cairo edition can be seen by considering the snippet of text "*shurakāʾa khalaqū ka-khalqihi*" from Sura *al-Raʿd* (13), verse 16, as given in the illustrations below:

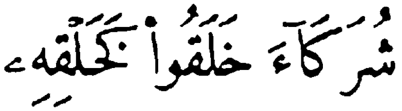


FIGURE 1.1A Example from Cairo edition

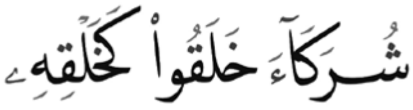


FIGURE 1.1B Example from Muscat edition

We only need to consider the last word group, *ka-khalqihi*, "like His creation." Notice how the angle of the first letter, the *kāf*, is significantly blunter than the *kāf* in the first word, in the Cairo edition. This is, it seems, only so that the *kāf* remains within the 'rails' of the typesetting. We also see that the dot of the second letter, the *khaʾ*, is placed to the right of the *kāf*, apparently for no other reason than that it was easier to typeset it that way. In Milo's version, the two *kāf*s have an identical angle and the dot is restored to its place to the left of the *kāf*, above the *khaʾ*. In other words, this digital rendering restores the correct

[190] For a history and evaluation of DecoType's work, see Nemeth, T. *Arabic Type-Making in the Machine Age: The Influence of Technology on the Form of Arabic Type, 1908–1993*. Leiden: Brill, 2017, pp. 410–434. Nemeth concludes that "the contribution of DecoType to Arabic type-making has been remarkable and its influence is here to stay."

way of writing Arabic, as attested in manuscripts, against the printed edition which saw itself limited by its own technology.

Beyond these improvements, Milo's software gives users the control to change the appearance of the text as far as the script allows. In this case, the space between the last two letters, the *qāf*, and the *haʾ*, can be elongated, the shape of the *kāf* can be changed, the two dots above the *qāf* can be arranged vertically, and lastly, the final *haʾ* can also get a floating miniature *haʾ*. The interactivity of the digital world is, thereby, optimally used, exactly to give back some of the fluidity and variety of the manuscript world. We see, then, an interesting reach from the digital world towards the manuscript world, one that can subvert the hegemony of the print world.

5 Consequences for Digitized Manuscripts

We have seen how the manuscript, print, and digital world can, at times, fold into each other and at times, go different ways. Their world views, their *epistemes*, are different. What happens when manuscripts and digital documents are meshed together? This could, of course, go in two different manners. One is that the contents of the manuscript are converted to digital format; the other is that the appearance of the manuscript is converted to digital format. The first option can be fairly well analyzed with the case studies just discussed, and we will look further into it in Chapter Five. In Chapter Two, we will consider the second option, which is the more pertinent one: when people speak of digitizing manuscripts, they mean taking digital photos of it. These photos could be of a full folio or part of it; they could be one or many, and they can be stored as they are or within a software environment that dictates the way we access them. The photos themselves can vary in quality, depending on the camera and the studio conditions, and also the post-processing. Lastly, using digital photos of manuscripts means we are operating in multiple worlds at once. Through the digital world, we engage with objects of the manuscript world. Moreover, this takes place in a time when the print world still looms large; for example, we often work towards creating a print publication. Given the vastly different world views these different worlds entail, working with manuscripts on a computer towards a print publication is a process in which we can make many false assumptions or fail to realize new opportunities. It is, then, time to seriously consider what it means to work with a digitized manuscript, in the next chapter.

CHAPTER 2

The Digital Materiality of Digitized Manuscripts

A strange paradox has crept into philological work of all kinds of fields, such as classics, sinology, Islamic studies, medieval studies, and many others. Scholars have been so eager to take advantage of digitized manuscripts, that Wido van Peursen, a biblical scholar, notes: "now that the digital object has become available, who will ever go back to the 'real' manuscript?"[1] This in itself may already be a paradox, suggesting that the more our research focusses on the manuscript world, the more our work takes place in the digital world. However, the paradox I have in mind is about what happens next. For, when it is time to disclose our sources, we refer to the actual, material manuscript and seem to forget we ever looked at digital images. Can we identify the digital surrogate so strongly with the material artifact? Or should we say that the two have nothing to do with each other? After a discussion of how digital resources are being incorporated in a paper-based scholarly discourse, I shall discuss these two positions. I shall conclude that both are untenable. Instead, for evaluating the use of digitized manuscripts we need to describe their 'digital materiality'. I introduce ten categories that can give shape to such a description and I end with a forceful call to include such a description in our publications, when we have made use of digitized manuscripts. Disclosing the use of a digital surrogate, instead of pretending like we accessed the material artifact, should become part of sound, scholarly conduct.

1 Stepping into the Digital World

It is generally understood that we are transitioning from a print world to a digital world, as explained in the previous chapter. Our own humanities fields are good examples of this. Publications in our subjects have a long shelf life and we therefore often require books and articles that appeared before the advent of digital publishing and only appeared in print. Yet, almost all currently active

[1] Peursen, W. van. "Text Comparison and Digital Creativity: An Introduction." pp. 1–30 in *Text Comparison and Digital Creativity*, edited by W. van Peursen, E.D. Thoutenhoofd, and A. van der Weel. Leiden: Brill, 2010, p. 10.

© CORNELIS VAN LIT, 2020 | DOI:10.1163/9789004400351_004

peer-reviewed journals publish their articles digitally.[2] The rather obvious answer to bridge the gap has been to digitize relevant print resources. Let us, before we proceed, arrive at precise definitions surrounding digitization. The terms 'digital' and 'digitized' are adjectives denoting the state an object is in. Digital print sources are printed works that are now digitally encoded as full-text, meaning that you can search through the contents, select it, and manipulate it (with limitations). We encounter them chiefly in two different flavors: either as plain text or as PDFs matching the lay-out and formatting as the print source. In the latter case, the file is almost certainly coming from the publisher. Digitized sources are photos or scans of pages of a print source. They can be a folder with as many image files as there are pages, or a PDF combining all images. The model used by platforms such as JSTOR, in which you see a scanned image of an article while still being able to select the text, is a combination of a digitized layer (the image) and a digital layer (the text). I shall use 'digital surrogate', 'digital photo', and 'digitized manuscript' more or less synonymously. By this I mean a collection of images of actual pages or page-spreads of a manuscript work. The qualification 'surrogate' is important, in order to emphasis the derivative nature of a photo which can only stand in for the original item in a limited manner, as will be amply discussed. Lastly, I frequently use 'digital repository', by which I mean a collection of either digital or digitized files, or a combination thereof, usually offered as one whole through a single website. In the literature, especially among librarians and archivists, it is often called a 'digital library', but I wish to forego a too careless identification of such digital resources with a brick and mortar library.

Four benefits of using digital repositories are generally mentioned. Using digital and digitized sources will speed things up, as one does not have to go to a library, collect, request, or recall all the necessary books and journals, and either photocopy them or keep extensive notes. It will, in addition, provide access to resources that are otherwise hard to get. Especially studying manuscripts benefit from digitization in this regard, as each is unique and can therefore only be accessed in one place on Earth, usually in a library with strict rules regarding their handling. Next to these two aspects of access on the item level, there is also a perceived benefit of being able to dig within a source directly to the desired passage. This is especially true for digitized sources of which the contents are fully searchable. Lastly, and more wishfully than actually so, it is

[2] Collins, E., and M. Jubb. "How Do Researchers in the Humanities Use Information Resources?" pp. 176–87 in *Liber Quarterly* 21, no. 2 (2012), p. 177.

said that digital repositories can open up new avenues of research, by allowing entirely new research questions, methods, and ways of publishing results.[3]

Consistently, however, it is print or manuscript sources that are cited by scholars, even if they used a digital surrogate. It does not matter if a scholars engages with a digital or digitized source, nor whether the source is originally an article, book, archival document, manuscript, or some other source from the manuscript or print world.[4] A number of reasons can be identified.

Print is immutable, whereas digital is mutable. Once ink has formed a shape on paper, it is exceedingly hard to forge that letter into the shape of another letter. To do this for a whole sequence is near impossible. Printed works principally derive from their printing press, where their pages have been typeset. Thus, each item is an exact replica of another item, and in itself only offers the privilege to be read. It itself cannot be used to create another copy; only the original printing press can. Digital works, however, can themselves be changed at will and be replicated and disseminated in that new state. Access to them necessarily implies reading and writing privileges. An extreme example are entries for Wikipedia, whose contents undergo a process of change from moment to moment as each user is not only allowed to read it but also to fundamentally change it. Citing and critiquing a digital source becomes fraud, as by the time a reader compares the digital source, it might have changed and it will look as though the author who is giving the critique is doing so unwarranted. Scholarship needs fixed points of reference, which print can give and digital cannot.

Closely related is the reason that a printed work is much more self-reliant than a digital work. No electricity is needed, nor an internet connection, operating system, plug-ins, etcetera. Furthermore, the physical devices to store digital data, such as CD-ROMs and hard drives, generally have a life span as low as a few years, which is laughably short compared to printed works that are still functional after hundreds of years. This makes it seem that a printed work is more robust and future proof. Indeed, it is not hard to come up with

[3] Collins, E., M.E. Bulger, and E.T. Meyer. "Discipline Matters: Technology Use in the Humanities." pp. 76–92 in *Arts & Humanities in Higher Education* 11, no. 1–2 (2011), p. 81; Rimmer, J., C. Warwick, A. Blandford, J. Gow, and G. Buchanan. "An Examination of the Physical and the Digital Qualities of Humanities Research." pp. 1374–1392 in *Information Processing and Management* 44 (2008), p. 1375; Nichols, S.G., and N.R. Altschul. "Digital Philology: A Journal of Medieval Cultures." pp. 1–2 in *Digital Philology: A Journal of Medieval Cultures* 1, no. 1 (2012), p. 1.

[4] Robinson, P. "Current Issues in Making Digital Editions of Medieval Texts—or, Do Electronic Scholarly Editions Have a Future?" *Digital Medievalist* 1 (2005); Collins, E., and M. Jubb. "How Do Researchers …", p. 182; Rimmer, J., et al., p. 1378; Collins, E., et al. "Discipline Matters", p. 81.

examples of resources once digitized and digitalized with great effort, which have already been rendered useless.[5] Scholarship needs stability in its data, for otherwise the research is not replicable. This is something print can give and digital does not.

This brings us to curation, which can be wanting for digital works. It seems that most digitization efforts, turning manuscript and print sources into a digital surrogate, are project based. This means that once the project is over, the money dries up and the collection is no longer maintained. Meanwhile, the digital environment is one for which maintenance is crucial, as hardware and software need to be upgraded regularly. Printed sources require less curation, and for the part that they do, they are in the comfortable environment of a library which takes a programmatic approach to maintaining its holdings.

Digitized works are not cheap. They are most often accessed on a subscription model, in which the user or library does not buy ownership of the digital file, but buys access to it. Commercial vendors have stepped into this market, which often enforce legal limits to usage far beyond the protection of printed works. The rights of a scholar to retain a private copy are purposely pushed into a grey area.[6] Libraries actively work against their users in this regard, for example by limiting the total number of scans someone is allowed to make. It is unclear how this will develop in the coming years. Indeed, this commercial and legal issue is highlighted by many scholars as one of the greatest challenges.[7] Print, on the other hand, has worked out these issues long ago and we know what to aspect from it.

Digitization, further, regularly gives the impression of being complete. However, those who dig deep to find a particular edition of a work purported to be included in the Google Books or Microsoft Digitization projects, will often find that they are out of luck. It is likely that only one edition of a printed work is digitized, which may still be of use in the initial research phase but becomes unusable in the writing phase when it is often times crucial to cite a

[5] Rimmer, J., et al., p. 1377.

[6] Besek., J.M., et al., "Digital Preservation and Copyright: An International Study," pp. 104–111 in *The International Journal of Digital Curation*, vol. 2, no. 3 (2008).

[7] Muri, A. "The Grub Street Project: Imagining Futures in Scholarly Editing." pp. 15–26 in *Online Humanities Scholarship: The Shape of Things to Come*, edited by J. McGann. Houston: Connexions, 2010; McGann, J. *A New Republic of Letters: Memory and Scholarship in the Age of Digital Reproduction*. Cambridge Mass.: Harvard University Press, 2014, pp. 20, 133; Rimmer, J., et al., p. 1387; Prescott, A. "Consumers, Creators or Commentators? Problems of Audience and Mission in the Digital Humanities." pp. 61–75 in *Arts & Humanities in Higher Education* 11, no. 1–2 (2011), p. 65.

first or most recent edition. The user is forced to abandon the digitized work in favor of the printed work.

Lastly, then, it also comes down to perceived authority. If a certain printed edition is a standard reference point, scholars are inclined to give a similar reference rather than admitting they used a digital surrogate of this standard edition. An important aspect of the diminished authority of digital surrogates is their accessibility. Even though digitized works are far more accessible, they are so in a siloed manner. As Jerome McGann, scholar of English studies, argues, such a digitized work "lacks the professional infrastructure that the scholarly book possesses by virtue of the mature social network in which it is located."[8] He goes on to say that:[9]

> Internet ecology at present is volatile and promiscuous, it encourages individual initiatives and "just-in-time" collaborations rather than programmatic strategies. This happens because internet culture has yet to map itself to the complex social system that powers scholarship and education.

McGann's observation has implications for both consumption and production of digital repositories. Let us focus here on the consumption, and conclude that the adage 'if you build it they will come' is suspect in the case of digital repositories, as it is unclear how prospective users would even know of it, nor how they can evaluate its quality.[10] In other words, digital repositories are not sufficiently linked to each other.[11] Moreover, to the extent that such links are created (for example within a repository, listing books that are related to the one the user is currently seeing), a digital context is perceived as unmanageable,[12] or, in other words, disorganized.[13] The individual items of a siloed repository thereby become siloed themselves. They are accessed either because the user knows the title beforehand, or because the item contains the keyword that the user was searching for. In comparison to an open stacks library, this precludes

[8] McGann, p. 30.

[9] McGann, p. 136.

[10] Cf. Warwick, C., M.M. Terras, P. Huntington, and N. Pappa. "If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data." pp. 85–102 in *Literary and Linguistic Computing* 23, no. 1 (2008).

[11] Collins, E., and M. Jubb. "How Do Researchers …", p. 185.

[12] Rimmer, J., et al., p. 1384.

[13] Jeanneney, J.-N. *Google and the Myth of Universal Knowledge: A View from Europe*. Translated by T.L. Fagan. Chicago: The University of Chicago Press, 2007, p. 67ff.

the serendipity of running into works one had not even imagined to look for, the so-called unknown unknowns.[14]

For humanities scholar working with manuscripts, the isolation of digital repositories can be a real problem. Peter Ainsworth, scholar of French studies, finds it a paradox that there is an "absence to date of any new, custom-built or standardised electronic tools to take over the role of the microfilm reader."[15] Likewise, Matthew Fisher, scholar of English studies, suggests that such a lack is at the basis for issues of trust and authority in using digital repositories. He imagines that tools and frameworks that can operate and interact between different repositories can solve this.[16] To make up for this lack and to get on with research, digital assets are approached as though print or manuscript assets. Ryan Szpiech, scholar of Romance languages and literatures, notes that "the academic study of medieval manuscripts has always been undertaken as if in dialogue with printed books",[17] and Adriaan van der Weel, scholar of book studies, adds that "most of our information habits remain book based."[18] And so, as some advantages of digitization are readily admitted, and digital and digitized materials are often used without thinking twice about it, there are also some sharp limitations put on its use, especially when moving from researching (using) to publishing (admitting to use).

2 Larger Than Life Digitized Manuscripts

The use of material manuscripts, meanwhile, has been substituted largely in favor of digital surrogates. The few scholars who have reflected on the impact of this change in our work have been emphasizing the larger than life quality of digital surrogates. By this I mean that with a digital surrogate, one can

[14] A concept that became publicly known after its use by then US Secretary of Defense Donald Rumsfeld, in a briefing on February 12, 2002. Cf. Morris, E. "The Certainty of Donald Rumsfeld." *The New York Times*. March 25, 2014.

[15] Ainsworth, P. "E-Science for Medievalists: Options, Challenges, Solutions and Opportunities." *Digital Humanities Quarterly* 3, no. 4 (2009).

[16] Fisher, M. "Authority, Interoperability, and Digital Medieval Scholarship." pp. 955–964 in *Literature Compass* 9, no. 12 (2012). The most promising framework, for our fields, is the *International Image Interoperability Framework*, to be discussed in Chapter Five.

[17] Szpiech, R. "Cracking the Code: Reflections on Manuscripts in the Age of Digital Books." pp. 75–100 in *Digital Philology: A Journal of Medieval Cultures* 3, no. 1 (2014), p. 75.

[18] Weel, A. van der. "New Mediums: New Perspectives on Knowledge Production." pp. 253–268 in *Text Comparison and Digital Creativity*, edited by W. van Peursen, E.D. Thoutenhoofd, and A. van der Weel. Leiden: Brill, 2010, p. 254.

supposedly "enhance an image's size",[19] or, as Maura Nolan, scholar of English literature, puts it, digital surrogates can reveal details otherwise "not visible to the naked eye."[20] Indeed, the applicability of the term 'surrogate' has been called into question, since the digital images "will often provide more information than would simple access to the physical object."[21] Thus, Ainsworth boasts of digital surrogates at 500 dpi (dots per inch) with each image file being around 150MB.[22] The West Semitic Research Project has captured digital photos of Dead Sea scrolls with a resolution so high one can inspect hair follicle patterns and surface deterioration, and made infrared photos to reveal the text more clearly.[23] The end result of this race for magnification is unsurprising; scholars have already gone as deep as the atomic level.[24]

Digitization on an atomic level should be seen as a reductio ad absurdum of the predilection towards high quality, challenging the whole concept of larger than life digitized manuscripts. It should make us ask how much larger than life digitized manuscripts really are. This has not happened, so far. Instead, digital surrogates are considered outright replacements of the original manuscript. Because of the enthusiasm surrounding the quality of digitized manuscripts, Nollan writes that "digitization […] offers a new and improved version of the original medieval object."[25] Szpiech argued that if scholars are under the impression that digitized manuscripts provide a better image than the material manuscript, we will undoubtedly prefer it and more or less forget about the original, material manuscript.[26] To a large degree, we have indeed, across the different fields working with manuscripts, forgotten about the material manuscript. Even the scholars who have reflected on the relation between material

[19] Rimmer, J., et al. "An Examination of the Physical …", p. 1375.

[20] Nolan, M. "Medieval Habit, Modern Sensation: Reading Manuscripts in the Digital Age." pp. 465–476 in *The Chaucer Review* 47, no. 4 (2013), pp. 470, 472.

[21] Crane, G., A. Babeu, D. Bamman, L. Cerrato, and R. Singhal. "Tools for Thinking: ePhilology and Cyberinfrastructure." pp. 16–26 in *Working Together or Apart: Promoting the Next Generation of Digital Scholarship*. Washington, D.C.: Council on Library and Information Resources, 2009, p. 23.

[22] Ainsworth.

[23] Hunt, L., M. Lundberg, and B. Zuckerman. "Concrete Abstractions: Ancient Texts as Artifacts and the Future of Their Documentation and Distribution in Their Digital Age." pp. 149–172 in *Text Comparison and Digital Creativity*, edited by W. van Peursen, E.D. Thoutenhoofd, and A. van der Weel. Leiden: Brill, 2010, p. 154.

[24] "Books Under the Microscope." *UT News: The University of Texas at Austin*, October 18, 2012; Treharne, E. "Fleshing out the Text: The Transcendent Manuscript in the Digital Age." pp. 465–478 in *Postmedieval: A Journal of Medieval Cultural Studies* 4, no. 4 (2013), p. 471.

[25] Nollan, p. 471.

[26] Szpiech, pp. 93–94.

manuscript and digital surrogate betray a casual identification between the two. One example will, I think, suffice. Ryan Szpiech relates his experience of finding a digitized manuscript on the internet in the following manner:[27]

> I follow the first link, and I am off to Portugal, to the library of the Universidade de Coimbra, where there is an extraordinary multilingual manuscript (MS 720) that has, so far, been little studied. With two more clicks I am viewing it in high resolution.

The impression one gets is that Szpiech is treating the screen of his computer as a real, physical window, through which he can look from Michigan where he works, onto Coimbra, specifically, the manuscript MS 720.[28] Granted, he does say he is looking at it "in high resolution," but this is merely the predilection for high quality at work. More important is that he says that he views "it", meaning the material manuscript, not the digital surrogate. Indeed, at the end of his article when he lists all his sources, he lists the digital surrogate that he looked at as "MS 720, Biblioteca da Universidade de Coimbra, Coimbra." This means that he is citing the material manuscript, which he confirmed he never actually saw (nor touched, smelled, tasted, or heard). Any trace of him having consulted the digitized manuscript is gone, and with it the ability for a critical reader to replicate his editorial and analytical decisions.

To understand why people would do this, I wish to invoke the philosopher Walter Benjamin:[29]

> Even the most perfect reproduction of a work of art is lacking in one element: its presence in time and space, its unique existence at the place where it happens to be. […] The presence of the original is the prerequisite to the concept of authenticity. […] [T]hat which withers in the age of mechanical reproduction is the aura of the work of art.

In this context, we can consider manuscripts to be works of art, making print publications the reproductions of which he speaks. The thinking laid out here has had an immense influence on humanities scholars. By editing a variety of manuscripts into one printed work, the aura of each manuscript became

[27] Szpiech, p. 76.

[28] A similar statement is made by Maura Nolan who writes "I can sit in Berkeley, California, and look at a manuscript in Oxford, England, thousands of miles away." Nollan, p. 473.

[29] Benjamin, W. "The Work of Art in the Age of Mechanical Reproduction." pp. 217–251 in *Illuminations: Essays and Reflections*, translated by H. Zohn. New York: Schocken Books, 1969, pp. 220–221.

lost, as the thinking goes.[30] This explains the enthusiasm of people for digitized manuscripts, as it brings back some of that aura. Especially when several copies of the same work are brought together, digitized manuscripts can restore the plurality that the manuscript world offered and which the print world took away, Francesco Stella, scholar of medieval Latin literature, argues.[31] As Farkas Gabor Kiss, scholar of renaissance literature, puts it, using digitized manuscripts can replicate the medieval reading experience.[32] Or, as Van Peursen puts it, this makes new realizations of 'presence' possible.[33] Or consider Deborah McGrady, scholar of French literature, who writes that "medieval manuscript readers and modern viewers of the digitized codex […] share a nostalgic desire for a multi-sensorial and intimate encounter with the textual body."[34] Especially this last comment by McGrady provides some explanation why reference is made to the material manuscript and not to the digital surrogate, as the aim would be to capture the supposed aura of the material manuscript. A pathological desire to reassert the aura is what Jacques Derrida has called 'archive fever': "a nostalgia for the return to the most archaic place of absolute commencement."[35] With academic libraries restricting access to their manuscript collections more and more, equally due to budgetary restrains as due to conservation concerns, digitizing manuscripts became the path of least resistance to let this fever run its course.

If you think that digitized manuscripts cannot possibly bring back the aura of the original, since it itself is a mechanical reproduction, you are not alone. Evyn Kropf, librarian at the University of Michigan who led an effort to digitize the entire collection of more than a thousand Islamic manuscripts, shows astonishment that some scholars have published on her manuscripts relying

[30] See Szpiech's use of 'authentic' to describe manuscripts; Szpiech, pp. 77–78.

[31] Stella, F. "Digital Philology, Medieval Texts, and the Corpus of Latin Rhythms, a Digital Edition of Music and Poems." pp. 223–249 in *Digital Philology and Medieval Texts*, edited by A. Ciula and F. Stella. Pisa: Pacini, 2006, p. 229. He is undoubtedly influenced by Dagenais, J. *The Ethics of Reading in Manuscript Culture: Glossing the Libro de Buen Amor*. Princeton: Princeton University Press, 1994.

[32] Kiss, F.G., and et al. "Old Light on New Media: Medieval Practices in the Digital Age." pp. 16–34 in *Digital Philology: A Journal of Medieval Cultures* 2, no. 1 (2013), p. 21.

[33] Peursen, W. van, p. 8. 'Presence' is the Modern Literature variant of what 'New Philology' is for Philology and the 'Material Turn' is for Anthropology, cf. Gumbrecht, H.U. *Production of Presence: What Meaning Cannot Convey*. Stanford: Stanford University Press, 2003.

[34] McGrady, D. "Textual Bodies, the Digital Surrogate, and Desire: Guillaume de Machaut's Judgment Cycle and His Protean Corpus." pp. 8–27 in *Digital Philology: A Journal of Medieval Cultures* 5, no. 1 (2016), p. 8.

[35] Derrida, J. *Archive Fever: A Freudian Impression*. Translated by E. Prenowitz. Chicago: The University of Chicago Press, 1996, p. 91.

entirely on the digital surrogate.[36] Says Kropf: “Consulting both original and surrogate seems appropriate, as there may be some qualities of the original that a surrogate does not mediate well.”[37] What could these qualities be?

3 The Intangible Aura of Material Manuscripts

Consider the following similar statements, apparently written independently:

> Taste, smell, and touch tax our ability to describe sensation in language and thereby to communicate the experience of handling a medieval book, which allows aura to maintain its status as a secret.[38]

And:

> The manuscript cannot only be seen—it must be touched, smelled, read, received, interpreted in order to be appreciated and understood. It can be appreciated fully only by means of a give-and-take relationship, and in that relationship, it will always remain partly elusive.[39]

And:

> Without a multi-sensual embodied experience of the material artifact, we experience only the transcendent, the partial; and we only ever grasp a fragmented and limiting understanding of the book’s intrinsic aura.[40]

These statements come from Maura Nolan, Ryan Szpiech, and Elaine Treharne,[41] and all speak of using a manuscript as a bi-directional contact that involves not

[36] Kropf, E.C., “Will that Surrogate Do?: Reflections on Material Manuscript Literacy in the Digital Environment from Islamic Manuscripts at the University of Michigan Library,” pp. 52–70 in *Manuscript Studies* vol. 1, no. 1 (2017): p. 67. Cf. Arnold, D. “Digital Artefacts: Possibilities and Purpose.” pp. 159–70 in *The Virtual Representation of the Past*, edited by M. Greengrass and L. Hughes. Farnham: Ashgate, 2008, p. 159; Correa, D.J. “Digitization: Does It Always Improve Access to Rare Books and Special Collections?” pp. 177–79 in *Digital Technology & Culture* 45, no. 4 (2017), p. 177.

[37] Kropf, p. 54.

[38] Nolan, p. 476.

[39] Szpiech, p. 90.

[40] Treharne, p. 477.

[41] There are more scholars who support this view, see: Peursen, W. van, p. 9; Rimmer et al., p. 1376; Terras, M.M. “Artefacts and Errors: Acknowledging Issues of Representation in

only all five senses but also our internal capacity to interpret and communicate the holistic experience that emerges from this contact. Further, they both emphasize that even then something of that aura that is created (or experienced) at contact remains unknown or ineffable.

Nolan seeks this ineffability in the magical. She concludes that "these quasi-religious aspects of manuscript study imbue books with the magic of the past."[42] It may feel that way, but not because of the manuscript. I wish to suggest that it can be largely a projection of our own experience onto the manuscript. Every manuscript was at one point new and could therefore not possibly invoke the past. Furthermore, I suspect that the 'magic' Nolan experiences is largely dependent on modern-day library conditions. Such manuscripts can only be seen by appointment, in a designated room in which you may only enter after stowing your bag. Pens are strictly prohibited. Manuscripts usually come in a cardboard box and need to be laid down on a pillow with lead snakes to keep the manuscript open. You are surrounded by silent people, each working on another ancient artifact. All these aspects make reading a manuscript a very special, unusual thing to do. And all these aspects are foreign to how people used manuscripts in centuries before. Pre-modern people had manuscripts in their own possession and used them intensively. One of the most normal things to do while actively reading was to write notes in the margin. Go ahead and try to write a gloss in the margin of a manuscript next time you are at the library; see how far you will get before librarians come rushing from every side. So in this sense what we experience with a manuscript is not essential to the manuscript itself, but created ('performed' if you will) by our current context.

Szpiech does not specifically address the elusiveness of manuscripts but he hints at an interpretation worked out more fully by Treharne. She makes an excellent point about the difficulty of judging the size, weight, use state, and material costs through inspection of a digital surrogate,[43] but it is another thing to say that "this hapticity is essential."[44] To argue for that, she writes that: "With a medieval book, the fortunate momentary owner touches, skin-upon-skin, in direct tactile intimacy with the very people who compiled and wrote those books."[45] In this interpretation, the contact between manuscript and reader is made out to be almost sensual, skins rubbing against each other

the Digital Imagining of Ancient Texts." pp. 43–61 in *Kodikologie Und Paläographie Im Digitalen Zeitalter 2*, edited by F. Fischer, Chr. Fritze, and G. Vogeler. Norderstedt: BoD, 2010, p. 55; McGrady, p. 9.

[42] Nolan, p. 475.

[43] Treharne, p. 476.

[44] Treharne, p. 471.

[45] Treharne, p. 474.

creating an intimate encounter. Nobody would disagree with Treharne when she says that "little of this intimacy and involvement, this prosthetic function, can be represented in the digitized image of a manuscript."[46] After all, only visual aspects are represented in a digital surrogate, and even then in a mediated manner, converting a three-dimensional perception to a two-dimensional one, possibly distorting the size and color of the original object.[47] Even when we use a touch screen we are always one step removed from the object.[48] Many scholars will remember the little moments of admiration when, for example, you find a smudge of a fingerprint in the margin of a medieval manuscript, or some ancient hair stuck in between pages. Equally, good materials like ink, paper, and leather can certainly be appreciated by codicologists. But these are tiny moments in between the real work and are accidental to it, not essential.

To be more precise, I have argued that using digitized manuscripts is a different experience from using material manuscripts in modern libraries, which is a different experience from how manuscripts were used historically. There are two conclusions to be drawn from this. Firstly, even though using digitized manuscripts will not get you the same experience as using a material manuscript, so is using a manuscript of hundreds of years old in a university library not the same as using a manuscript of just a few years old in your private possession. I therefore deny claims that this aural, ineffable experience of handling a real manuscript is "clearly paramount" (Treharne) or "essential" (Nolan). Secondly, it is not like using digitized manuscripts will get you none or only a shadow of the experience you would get from handling the real artifact. Rather, you simply get another experience. Since we established that such an experience is constituted by the contextual performance of handling a manuscript, each experience (all three) are equally valid.

I therefore conclude that discussions on the relationship between digitized and material manuscripts have brought us little to nothing. It is neither correct to see digitized manuscripts as better than their material counterparts, nor is it correct to see them as worse than them. The root problem for both mistaken interpretations is that the 'digital materiality' of digitized manuscript is not

[46] Treharne, p. 474.

[47] This reduced experience has upset some people so much that they spent considerable efforts to salvage it, see e.g. Chu, Y., D. Bainbridge, M. Jones, and I. Witten. "Realistic Books: A Bizarre Homage to an Obsolete Medium?" pp. 78–86 in *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, New York: ACM, 2004.

[48] Mangen, A. "Hypertext Fiction Reading: Haptics and Immersion." pp. 404–19 in *Journal of Research in Reading* 31, no. 4 (2008), p. 405.

understood.[49] We would only need to consider how extraordinarily specific technical guidelines for capturing manuscript materials and print publications are,[50] to make us realize not all digital surrogates are created equal. Such standards are necessary, since, as Mats Dahlström sums up:[51]

> digitisation and the subsequent editing of images has perhaps more than any other editing phase made us attentive to the fact that virtually all parameters in the process (image size, colour, granularity, bleed-through, contrast, layers, resolution etc.) require intellectual, critical choices, interpretation, and manipulation.

If the producers of digital surrogates are aware of the influence technical decisions have on the outcome, so should we as consumers. How do we assess the quality of digital surrogates? The only scholar who has identified this question, that I am aware of, is Melissa Terras, scholar of digital humanities. In search for an answer she is a bit helpless. She timidly concludes that "it can be difficult to assess the quality of digital images," and puts the onus back on the producers of digital surrogates, arguing that "it is imperative that those undertaking digitisation programs consult guidelines and carry out benchmarking procedures to ensure quality control of the digitised output."[52] This, to me, is unacceptable. We cannot hold our libraries accountable for mistakes we make based on digital surrogates. Besides, the discussion on how to properly digitize has already played out for the most part, and with hundreds of thousands of manuscripts already digitized there is little we can do in influencing their quality. Instead, we should simply acknowledge the digital materiality of digitized manuscripts, appreciate the added value they bring, and work around their drawbacks. This, we cannot do as long as we are incapable of describing their digital materiality.

[49] For a similar discussion on art and museum pieces, see Cameron, F. "Beyond the Cult of the Replicant: Museums and Historical Digital Objects—Traditional Concerns, New Discourses." pp. 49–75 in *Theorizing Digital Cultural Heritage: A Critical Discourse*, edited by F. Cameron and S. Kenderdine. Cambridge Mass.: The MIT Press, 2007.

[50] See e.g. "Technical Guidelines for Digitizing Cultural Heritage Materials." Federal Agencies Digital Guidelines Initiative, 2016; Dormolen, H. van. *Richtlijnen Preservation Imaging Metamorfoze*. Den Haag: Koninklijke Bibliotheek, 2012.

[51] Dahlström, M. "Critical Editing and Critical Digitisation." pp. 79–98 in *Text Comparison and Digital Creativity*, edited by W. van Peursen, E.D. Thoutenhoofd, and A. van der Weel. Leiden: Brill, 2010, p. 81.

[52] Terras, p. 52. The difficulty of assessment had been noted in circles of production, cf. Lynch, C. "Authenticity and Integrity in the Digital Environment: An Exploratory Analysis of the Central Role of Trust." pp. 32–50 in *Authenticity in a Digital Environment*. Washington: Council on Library and Information Resources, 2000, p. 35.

4 What Are Digitized Manuscripts?

I have come to see this question as relating to two factors: the digital data that constitute the actual, digitized manuscript, and the digital context in which this is stored or offered. The first is one or more files, for example a folder full of image files, or a PDF with a photo on each page, or a combination thereof, for example when different qualities are offered. The second can take on many shapes but I shall call it in all cases a repository. This context can be as rudimentary as a collection of folders on your own computer, in which case we should look at the folder structure and file naming to evaluate the repository. It can also be as advanced as a web server hosted by a library, that dynamically serves up manuscripts through an interface that is enriched with catalog data. Ten notions can help us to analyze those two factors: (1) size of the collection; (2) online availability; (3) ability to download; (4) the portal; (5) the viewer; (6) indication of page numbers; (7) image resolution; (8) color balance; (9) lighting; and 10) how the image is cut.

The size of the digital repository (relative to the number of material manuscripts held) is an important way to assess the commitment a library has to digitization and understand the trajectory of digitization globally. Whereas in 2011 libraries with Islamic manuscripts would boast of two digit numbers of digitized manuscripts,[53] the most important ones have now scaled to four digit numbers and beyond. In a good number of cases, libraries adopted a strategy of blanket digitization, simply digitizing a whole collection without picking out the supposed important manuscripts and leaving out supposedly unimportant ones. Other libraries hold the policy that the first request pays. Whoever asks for a holding to be digitized will pay a fee for it, subsequently it can be accessed for free by anyone else. This ought to be preferable over another popular policy, namely to digitize as a project based on external funding. I have observed that several repositories started showing decay after this funding period was over. We shall see examples of this in the next chapter. In this regard, size is important for another reason, namely the hope that larger repositories will be more likely to receive continued funding, maintenance, and upgrading.

Online availability is, I think, quite crucial for any philologist, as it takes out expensive and time consuming travel to faraway libraries. However, users need to be aware that not everything can be instantly seen online. In cases in which digitized manuscripts are not offered on a website, their files need to be requested and usually the files will still be send over the internet. In other

[53] Cf. Swanick, S., "Of making books there is no end: Islamic manuscripts on the Web," pp. 416–419 in *College and Research Libraries News* (July/August 2011).

cases, digitized manuscripts are behind paywalls. This means you cannot see anything unless you pay first, after which you have online access to all the files. Of the twenty repositories I compare in the next chapter, the vast majority allow free, online access.

Such access does not always translate in the **ability to download**. Some repositories only allow online viewing and have implemented technological hurdles to impede users from downloading. Other repositories will add a watermark to every picture, making some readings more difficult and browsing more annoying. Others, again, only allow users to download one page at a time. Only about a third of the repositories I looked into allow full download. I think it is important to keep private copies of the manuscripts you work with on your own computer. This will ensure you can access them again, and will allow you flexibility in annotating them while you are working. In addition, if you intend to refer to manuscripts in publications, you may be asked by readers to provide your evidence so that they can double-check your analysis, which will be easier to do if you have the files in your possession. Related to this aspect of downloading is the copyright asserted over digital surrogates. This is discussed in further detail in the next chapter.

Another aspect to consider is **the portal**, which is the website which the user first enters, in order to find from there the actual photos of a manuscript. It is usually related to the manuscript catalog which a library has. In my own field of Islamic studies, the portal is usually very simple. This is definitely an area in which repositories can be expected to improve, allowing for new ways of browsing that go beyond a traditional catalog.[54]

The viewer is the technology used to allow users to see and browse the digitized manuscript. For close reading I find it better to download the manuscript and inspect it on my own. For quick browsing and skimming, however, the online viewer is a welcome technology. Here, too, improvement may be expected. Currently, most viewers are somewhat clunky, especially in browsing from page to page. Using technology makes sense insofar as it is helping us. When it works against us, which is the case for the majority of viewers in Islamic studies, we need to proceed with caution.

Page numbers (or folio numbers), together with the manuscript call number, are the basic descriptions that a user needs to make a reference to a

[54] Cf. Chevallier, P., L. Rioust, and L. Bouvier-Ajam. "Consultation of Manuscripts Online: A Qualitative Study of Three Potential User Categories." *Digital Medievalist* 8 (2013); Ornato, E. "La Numérisation Du Patrimoine Livresque Médiéval : Avancée Décisive Ou Miroir Aux Alouettes ?" pp. 85–115 in *Kodikologie Und Paläographie Im Digitalen Zeitalter 2*, edited by F. Fischer, Chr. Fritze, and G. Vogeler. Norderstedt: BoD, 2010, p. 96.

manuscript. In many cases, the viewer gives a wrong page number or none at all. Thus, the user is left to look at the folio numbers in pencil, as far as it exists in a manuscript and as far as it can be seen on a photo. It would be great if upon downloading the manuscript call number and page number could automatically be embedded in the file, for example hard coded in the bottom left corner and in the file name. In lieu of this, I would advise users to make sure to rename any downloaded file according to the provenance, including library or city name, manuscript number, and folio number. You may also find it helpful to rename the folder in which the file reside to indicate title, author, and manuscript number.

The image resolution is the most important metric in determining the quality and usability of a digital surrogate. Technically we are looking for the DPI or PPI: dots/pixels per inch, which tells you how many pixels were spent on storing one square inch of material manuscript. The higher the better, obviously, although it can take proportions that are unmanageable for private use. For example, the higher it gets, the higher the file size, which complicates downloading and storing. Also, very large images cannot be opened with any image viewer; simple ones will crash. The problem with DPI is that it cannot always be ascertained, or rather, it should always be doubted when not coming straight from your own camera. This is because images can and likely will be processed before they reach you, meaning that some software touched it and perhaps compressed it, reformatted it, or possibly enlarged it (without, obviously, any gain in quality). In this process, the metadata for DPI might have been altered or deleted. DPI can still be a useful measure when filing a specific digitization request with libraries or museums. Advice on this depends on your needs, but it is good to know that 300 is what is used in print media (magazines, newspapers and the like).

For any images we encounter, we do well to evaluate images with a combination of three indicators: dimensions, file size, and a visual impression. For example, the digital surrogate for Landberg MSS 711 folio 150a, at Beinecke Library, is said to be available in “high resolution.” Its downloadable JPG file is 2588 × 3415 pixels. Had it been stored in a format called “lossless JPG,” a variant of the JPG format that preserves the quality of the image as much as possible, this image should weigh around 13.3MB,[55] however, the size of the downloadable file is actually 2,3MB, more than eighty percent smaller than expected. In

[55] JPG file size is different from image to image, as even the lossless format performs file size compression techniques, see Murray, J.D., and W. VanRyper. *Encyclopedia of Graphic File Formats*. 2nd ed. Bonn: O’Reilly & Associates, 1996, pp. 191–205.

other words, some compression was applied to the image. This proves that not only image dimensions should be used to get a metric for the image quality. I should point out, though, that Beinecke in fact supplies fairly high quality images, as a visual inspection will quickly tell. Take for example images from McGill. The digital surrogate for Osler MS 389/23, folio 1b measures 4396 × 6116 pixels. It is therefore much larger in dimensions than Beinecke's image, yet the image from McGill only has a file size of 1,1MB, much smaller than Yale. Even so, McGill's images are usable. Take the images from the Parliamentary Library (Majlis) in Tehran as an example. Images of MS 5808 are approximately 650 × 1270 and have a file size of approximately 370KB. For reading purposes, this is not good enough. The resolution is so low, that here we should remind ourselves of Terras's fear that we cannot trust the readings from images that do not help us but rather work against us, in deciphering the manuscript.

In addition, a visual assessment is necessary. For example, Leiden University has given Brill publishers the rights to exploit digital surrogates of Islamic manuscripts. From the file size and dimensions one would not expect anything strange, but a visual assessment reveals that there is a peculiar jaggedness to the text, which severely reduces the quality in comparison to excellent resolutions such as the BnF offers through Gallica. Most of the repositories considered in the next chapter, however, offer resolutions that we can work with. Only in rare cases do we encounter digital surrogates for which the use for philological purposes is highly suspect.

The usability is not only influenced by the resolution. Also **color balance** plays a role.[56] What is meant by this is how true to color the digital surrogates are compared to the original object. This is hard to judge when we do not have the material manuscript to compare it with, but some tricks might give us a clue. If a repository offers different qualities of the same manuscript we can compare these different qualities.[57] If we have photos of more than one manuscript we can compare them. If they are strangely different, we know that the color balance is off for at least one of these photos. We can also look for

[56] Says Julia Craig-McFeely, historian of medieval music: "colour is one of the most devastatingly misleading fields in the digital imaging world." Cf. Craig-McFeely, J. "Finding What You Need, and Knowing What You Can Find: Digital Tools for Palaeographers in Musicology and Beyond." pp. 307–39 in *Kodikologie Und Paläographie Im Digitalen Zeitalter 2*, edited by F. Fischer, Chr. Fritze, and G. Vogeler. Norderstedt: BoD, 2010, p. 312.

[57] For example, from Beinecke Library I got two different files for MS Landberg 335, folio 1b–2a, one at 271kb with dimensions of 1200 x 908, the other at 1.6mb with dimensions of 3463 × 2620. The small file shows distinctly more red, given the page an earth tone, while the big file shows the page as more yellowish color.

common mistakes. These include that the paper looks green or blue or where we expect black and red ink we may see the ink color shift. Other issues are included in color balance as well, most notable the issue when the text of the back side of the folio is bleeding through on a photo. Especially this last issue can impact readability significantly. Color balance might simply interfere as a factor of annoyance. This is, I think, not to be brushed aside. When performing a close reading, you wish there to be as little distractions as possible and a severely skewed color can be such a distraction.

Much more of a distraction is **the lighting**. To make a photo, there obviously needs to be a light source. Ideally, the studio is brightly and evenly lit, giving the manuscript an even, bright lighting, resulting in a vibrant, sharp photo. This is not always the messy reality that we will encounter. Sometimes there is uneven lighting across different pages, giving a restless impression while browsing, as one needs to adapt to the different light intensity of each page. Sometimes valleys of shadow and hills of over-exposure are present. This means that parts of one page turn darker, while other parts turn brighter in the image without this being true for the material manuscript. This can happen when the page was not fully flat when the photo was taken, with the bulge of the paper creating these shadows, and it is a significant factor influencing the ability to read the text. Shadow and over-exposure can also occur when the photographers did not use a professional light source, with light coming clearly from a specific angle. In this case, especially, gilded parts of the manuscript can light up to the extend that they become unreadable. Something similar happens when a manuscript uses glossy paper, as some areas can bounce back so much light as to white out text in those parts. This happens especially at the top of the bulge of a page, if not flattened. Lastly, as far as I have come across, when manuscripts have been repaired using tape, a direct light source will invariably cause those parts to white out. In short, then, lighting can hamper readability (and hence, reliability) of a digitized manuscript and is a major cause for restless, uneven experience of the *mise-en-page*.

The last aspect that I found useful to consider is what I call **the cut**. This refers to the decision as to what is part of the photo and what falls just outside of it. A great cut means that the photo reveals a little space around the edge of the page and the cover, so as to assure a user that no visual aspect of the manuscript is hidden. Some repositories have cut their photos too tight. Their decision was to cut around the text block, leaving the edges of the page unaccounted for. For philologists this can be problematic when marginalia are important to consider, as there is the danger that parts of the marginalia are invisible. I am happy to notice that for the vast majority of digital Islamic manuscripts, the cut is not too tight.

5 New Habits for Philologists in a Digital World

Digitized manuscripts are a godsend for philologists, as they solve a problem we newly encounter. In the digital world we now work in, we usually approach pre-modern texts through the funnel of the print world, as it is printed materials that are converted to digital texts. Furthermore, whereas in the print world we frequently have the option of several editions, some of them with critical apparatus, in the digital world this is usually lost as only one edition is converted to digital text, usually without the apparatus, and even if multiple editions are available, we do not have convenient means to automatically iterate over all of them to spot meaningful differences. For distant reading, for example trawling big text corpora for finding larger trends, this does not need to be a great problem. However, if what we want is a close reading and we want to truly understand what an author is saying, this becomes a great issue. Being able to go back to the material evidence that manuscripts offer, by means of a digital surrogate, is therefore something that we will likely make use of more and more, to fully understand the text we are investigating.

If you are a philologist yourself you will probably need no convincing of this, but here is one example for those who are not entirely sure if this is such a problem. Take the legal ruling that the medieval theologian Ibn Taymiyya gave on whether Muslims could live in the city of Mardin (in the East of present-day Turkey) or should emigrate elsewhere since the city was not under Muslim rule. At first his answer seems nuanced and subtle, arguing for a compromise solution in which emigration is favored but not obligatory. But then the conclusion of the ruling puts all of this in an entirely different light: "[Mardin] is of a third type in which the Muslim should be treated as he merits, but the one outside of the Law (*sharīʿa*) of Islam should be combatted as he merits."[58] The different verbs, "be treated" (*yuʿāmal*) and "be combatted" (*yuqātal*), make an especial impact since the two sentences are constructed exactly the same. Like this it was printed in 1909 and from then on reprinted up until today. It has been quoted in this form by many leaders of terrorist groups,[59] and has found its way online and in a scholarly digital text corpus.[60] And yet it was pointed out by the Mauritanian scholar Abdullah bin Bayah that this should be

[58] Yahya Michot offers a slightly different translation: "Rather, [Mardin] constitutes a third type [of domain], in which the Muslim shall be treated as he merits, and in which the one who departs from the Way/Law of Islam shall be combated as he merits." See Michot, Y. *Muslims under Non-Muslim Rule*. Oxford: Interface Publications, 2006, p. 65.

[59] See Michot, *Muslims under Non-Muslim Rule*, pp. 101ff.

[60] I mean the Open Islamic Texts Initiative, which relies for this text on the digital text corpus *al-Maktaba al-Shāmila*. See Miller, M.T., M.G. Romanov, and S.B. Savant. "Digitizing

considered a misprint and that we should read “be treated” (*yuʿāmal*) in both cases.[61] Among the evidence he brings forth is one of the earliest manuscripts of Ibn Taymiyya’s text, which indeed shows *yuʿāmal*, not *yuqātal*.[62] It is evident, then, that establishing what the text is can be of literal vital importance.

The barrier to cross-check digital documents and print publications with manuscripts is getting lower and lower, as more manuscripts are available through digital surrogates. No doubt, we will therefore see a resurgence of manuscript studies. But as much as the higher usage of manuscript evidence is to be celebrated, we should be aware that digital surrogates cannot be used indiscriminately. Any digital resource of interest should be evaluated, which we can do according to the ten aspects that I outlined above: (1) size of the collection; (2) online availability; (3) ability to download; (4) the portal; (5) the viewer; (6) indication of page numbers; (7) image resolution; (8) color balance; (9) lighting; and (10) how the image is cut. Evaluating these aspects can answer the question to what extent the technology is working for you, rather than against you. This should be our leading principle in implementing digital photos in our philological work. The unfortunate result is that many resources cannot be used for any and all purpose. Many low resolution digital scans of microfilms exist. These are surrogates of surrogates. They can still be (and are) profitably used, for example to corroborate a particular reading. I am however skeptical of using them as a single source for making an edition. Perhaps, indeed, 99% of a manuscript can still be deciphered by using them, but it is about that 1% of cases in which the scribe fumbled a bit with his pen and it is unclear what the word reads. In those 1% cases, you do not wish to have a low-resolution, black and white reproduction of a reproduction as your sole witness. Even if you are steeped in the subject and can make an educated guess towards the correct reading, I think it is simply the duty of a scholar to concede that this is not a reliable way of working.

The effect stacks; if we have bad digital photos, we end up with bad editions, leading to bad translations, resulting in bad analyses. For many fields in the humanities, manuscripts remain the foundation of our work. So let us make this foundation as strong as we can. Our resources are limited and the number of manuscripts to be potentially digitized is incredibly large, so we can realistically only digitize things once. We have one chance to do it right,

the Textual Heritage of the Premodern Islamicate World: Principles and Plans.” pp. 103–109 in *International Journal of Middle East Studies* 50, no. 1 (2018).

[61] Michot, Y. “Ibn Taymiyya’s ‘New Mardin Fatwa’. Is Genetically Modified Islam (GMI) Carcinogenic?” pp. 130–181 in *The Muslim World* 101, no. 2 (2011), p. 145.

[62] Michot, “Ibn Taymiyya’s ‘New Mardin Fatwa’,” p. 146.

with an eye towards long term preservation and perhaps innovations such as automatically recognizing text within manuscripts. For resources yet to be digitized, this is something we ought to give our input on to the digitization experts in libraries and museums. For resources that are already digitized, the same principles applies: we can realistically only once critically edit a work, so let us get it right that first time. This means we should not use faulty digital surrogates that do not enable us to do our work but in fact hinder it. If that is all we have, we should rather chose to not edit it at this point in time than to go ahead and make our edition rely on poor digital surrogates. You may be thinking right now: "I am not aware of any published documentary material that has been read erroneously (and published, and refuted) due to faults in the digitization process."[63] This I find a maliciously ignorant stance towards the digital materiality of digitized manuscripts. We know that digitization can be of very high quality, but in the vast majority of cases is far from it. We also noticed that Ryan Szpiech did not disclose in his bibliography that he used the digital surrogate, instead citing the material manuscript as his source, and he is hardly alone in this practice. Misreads because of faults in the digital materiality of the surrogate are bound to happen, but if we do not tell our readers we made use of a digital surrogate, nobody will be able to call us out on it.

If, however, the quality of the photos checks out, I do not see serious objections to using exclusively digital surrogates in establishing a critical edition. For the classical work of editing a main text, concentrating on the text block, photos are frequently sufficient to establish the correct reading. It is, however, only fair that we do not let readers make the mistaken assumption we had access to the material manuscript. Being honest about this means that we should refer to the surrogate in our bibliography and we should include a description of the digital materiality of the photos, for example in the codicological description of the manuscript. This description need not be long, it can in fact be one sentence. In my own work I have used a sentence like this:

> *The digital images I work with are 2051 × 1925 showing two pages, at about 500kb. The cut is tight and color balance is fine.*[64]

This sentence conveys that I was using digital surrogates rather than the original manuscript. Then I described the general level of detail by combining the image dimensions and the file size, and mentioning whether one or two pages

[63] Terras, p. 56.

[64] Lit, L.W.C. van. *The World of Image in Islamic Philosophy*. Edinburgh: Edinburgh University Press, 2017, p. 257.

are visible. I then specified if the borders are visible or if information on the edge of the manuscript could fall outside the photos. I then mentioned if the photo is true to the actual color. These bits of information together give a sense of how legible a digitized manuscript is, and how professional the digitization was done. This, in turn, gives the reader a basis to judge how accurate the edition can be. In this case, the reader may notice that the images are not that great. It could be that some information in the margins of the manuscript is missing or not entirely legible due to the cut, and the file size is a bit small, as in my experience 1MB per page spread is generally a more reasonable file size. You can add to this description a note whether you had access to the material manuscript, and if so, what the purpose was of handling the manuscript itself, for example, in order to verify certain ambiguous readings. A description of the repository can also be included, based on aspects one through six, depending on how unusual these aspects are.

With that, we have come to a close of this chapter. Clearly, digital surrogates serve a very important function in our current workflow and are likely only to grow in their importance. Using digitized manuscripts can help us reach a more authentic reading of a text in a faster and cheaper fashion and has the potential to answer entirely new research questions. How problematic it is that this reduces our engagement to the sense of sight, at the cost of losing the rich mix of the other four senses, is something that we can evaluate on a case by case basis. Neither should we think that digitized manuscripts are a neutral window onto the material manuscript and we would be therefore fooling our readers were we to cite the material manuscript when we in fact consulted its digital surrogate. We can get a better grip on the usability of a digitized manuscript by evaluating it according to the ten categories introduced above and, in fact, we can use these to include a digital codicological description in our publications. A leading question we should ask is: is the digital surrogate helping me or hindering me, in doing my work? If the latter, we might have to make the hard decision of not going through with our work. What kind of variety in digitization you might reasonably expect to encounter is the topic of the next chapter.

CHAPTER 3

Digitized Manuscripts and Their Repositories, an Ethnography

In this chapter, I wish to chart the state that digitization is in, for the humanities worldwide, by evaluating twenty repositories. By doing so, the reader will learn more about what exactly makes digital manuscripts distinctly digital, how they are benefitting and sometimes obstructing the user, and how digital manuscripts are to be acquired. I shall close with an eye towards the future, exploring the strengths, weaknesses, opportunities, and threats of digitization of manuscripts.

This ethnography draws from my own field, Islamic studies, since that is what I can speak about with the most certainty. However, many repositories discussed here also hold non-Islamic manuscripts and for those which do not they will nonetheless shed light on the variety of digital surrogates you can encounter. Further, this ethnography was made in the summer of 2017. Some of these repositories have since then changed and that is only to be expected in a digital world. Even though important changes have been noted in footnotes, the assessment and the analysis have remained as they are, to be as fair as possible to all repositories. The repositories were selected for the different choices they made in digitizing, storing, and permitting access. Secondary attention was paid to the uniqueness of the libraries' holdings and their geographical location. The following is therefore not an exhaustive list of the world's best repositories, but rather a representative sample.

Five are from Europe: the State Library in Berlin, Leiden University Library, the British Library in London, the French National Library in Paris, and the Vatican Library.

Five are from North-America: the University of California, Los Angeles Library, Beinecke Library at Yale University in New Haven, Princeton University Library, University of Michigan Library in Ann Arbor, and McGill Library in Montreal.

Five are major digital collections from the MENA region: Suleymaniye Library in Istanbul, Topkapi Palace Museum Library in Istanbul, Malek National Library in Tehran, National Library of Morocco in Rabat, and King Saud University in Riyadh.

The final five are smaller or from other regions: Najah University in Nablus, Jafet Library at the American University in Beirut, Malaya University in

© CORNELIS VAN LIT, 2020 | DOI:10.1163/9789004400351_005

Malaysia, the Aboubacar Bin Said and Mamma Haidara Libraries in Timbuktu, and the Institute of Oriental Culture, University of Tokyo.

In the discussion on each repository I shall give a short description and explain how manuscripts are made available digitally. After individual descriptions, I compare the repositories under the ten notions introduced in the previous chapter. These notions can roughly be divided into two groups, one expressing the means of distribution, the other assessing the quality of the photos. To round out this discussion, I will finish this section with a comparison of the word *qāla* ('he said') as it appears in manuscripts from each collection. By keeping as many variables as possible constant, the appearance of this word should give us a visual overview of the range of quality currently on offer.

1 Old Collections in Europe

Libraries around Western Europe have been acquiring Islamic manuscripts for many centuries. An important characteristic of these libraries is that often times collectors chose their purchases well, bringing back some of the finest, rarest, and generally most interesting manuscripts back to Europe. Digital access to these collections is therefore crucially important.

1.1 *Staatsbibliothek zu Berlin*

The 'Stabi' in Berlin has to date more than 1,600 manuscripts digitally available online. Some of them come from digitization projects. Others were digitized upon request, following the first-customer-pays principle, in which the person first interested in a manuscript pays for it to be digitized, after which it becomes freely accessible online. The digital surrogates are integrated into their online catalog, which is therefore the principal entry, though through a special page one can be sure to find only digitized texts. The online catalog can behave erratic, giving different results for searches in Arabic, in simple transliteration, and in proper transliteration. It is therefore beneficial to also consult the printed catalog, which can be obtained in PDF format. From a catalog entry, one can click to open a viewer. This has most of the functionalities for efficiently viewing, browsing, and downloading the digitized manuscript.

1.2 *Qatar Digital Library*

The online digital holdings of the British Library relevant to Islamic studies is only a fraction of the actual collection, consisting at time of writing of only a bit more than 100 manuscripts. These are consolidated on a dedicated website, hosted in Qatar. Here one can find other types of sources as well, such as archival records, newspapers, and photographs. It seems to have been done on a project basis, with a topical concentration on the Gulf region. The website is sleek and clean and provides several ways to find manuscripts. It is advisable to use both Arabic and Latinized key terms when searching for specific words, as some entries are unevenly cataloged. Clicking on a record immediately takes you to images of the object. From a technical point of view, every image is a separate item and all images of one manuscript are linked. This means in practice that moving from one folio to another is time consuming. An ameliorating factor of this repository is that they included in the portal a page with 'Articles from Our Experts.' These are small case studies making use of the digitized materials, which are both informative and exemplary.

1.3 *Leiden Universiteitsbibliotheek*

The famous collection of Islamic manuscripts at Leiden is not freely available, nor fully. The university sold the right to Brill publishers, to commercially exploit the digital distribution rights of the manuscripts. A day pass is available for individuals. Brill decided to sell digitized manuscripts in batches, so far totaling less than 700. The portal seems to have gotten little thought. There is a long list of all manuscripts, without much description. Then there is a search function which did not perform well when I tested it. The best way to make use of this resource is to go through Jan Just Witkam's catalog to find a manuscript of interest,[1] and then use different search strategies to ensure the manuscript is included in Brill's modules. The viewer is decent, with options such as rotate, zoom, and download. A big drawback is that it is not indicated which folio you are looking at, or in fact which manuscript it is you are seeing. Given that the collection is behind a paywall, it is odd that the download function gives you inferior quality compared to zooming in, in the browser.

1.4 *Bibliothèque nationale de France*

The BnF has been in the business of digitizing for a long time, through a dedicated portal called Gallica. Islamic manuscripts can be found here too. There are seemingly in excess of 4,000 relevant items. Unfortunately, a large number of the digital surrogates are surrogates of surrogates: they are digitized microfilms. Digitizing microfilms is relatively inexpensive and fast, making it easier to establish a sizable digital repository. However, being two degrees away from the actual object impacts the usability. If, on the other hand, your manuscript of choice is available in full color, you are in luck, as Gallica has a

[1] Witkam, J.J. *Inventory of the Oriental Manuscripts in Leiden University Library*. 28 vols. Leiden: Ter Lugt Press, 2007.

superb portal and includes a superb viewer. The search function performs well, and the viewer is very flexible in how you want to browse, look, and download the manuscript.

1.5 *Biblioteca Apostolica Vaticana*

On the website of the Vatican Library, about 350 Islamic manuscripts are available, a fair number of them being digitized microfilms. They have only recently started a digitization project and intend to digitize their entire collection and make it freely available. Currently, you can only find what you are looking for if you know the exact manuscript you want. The website is arranged by collections and in each collection only manuscript numbers are given. Once you click on that, you see the images of the manuscript. The viewer has most of the desired features, including a precise indication which manuscript and folio you are looking at. Downloading can only be done one page at a time, and it includes a watermark of the Vatican.

2 New Collections in North America

It is estimated that about 33,000 Islamic manuscripts are to be found in North America, virtually all of them at universities.[2] Given the financial and technological means available in North America, it should be possible to have all manuscripts digitized, with a union catalog to hold it together. Most libraries are hard at work to make that a reality, and as such the libraries listed here are among the leaders in blazing a trail in digital preservation of Islamic manuscripts.

2.1 *Caro Minasian Collection at the University of California, Los Angeles Library*

I am including UCLA's repository for its unwieldy approach and its reliance on outdated technology. Less than 600 manuscripts have been digitized, as a capita selecta from the Caro Minasian collection. They are trapped in a catalog with fuzzy metadata. For example, there are six categories to tag a work as 'Philosophy', each with different works tagged. It also includes duplicates, e.g. search for "رسالة في الحكمة" and you will see two identical entries called *Risālah fī al-ḥikmah*. Incidentally, when one search for the term in Latin alphabet, "Risālah fī al-ḥikmah", it gives back zero results. The viewer is built with Adobe Flash, which has become obsolete and makes it impossible to download

[2] Some of them are in museums, such as The Walters and the Freer and Sackler Galleries.

images apart from screen-grabbing them.[3] It seems the work was done in the period 2007–2009 as a project, partly funded through an NEH grant. Despite promises of making it much larger, it has not been expanded, improved, or migrated to newer technology since then.

2.2 *Beinecke Rare Book & Manuscript Library*

The Beinecke Library, part of Yale University, has a state of the art digitization studio. It has placed its Islamic manuscripts online within the wider digital collections that they curate. This makes it unclear exactly what they offer freely online, though currently it seems to total less than 300 manuscripts, some of them only partially digitized. Searching can only be done using transliteration. The viewer is spartan and a bit of a hassle to work with, though it includes the option to download a page or the entire document.

2.3 *Princeton University Library*

At Princeton, two repositories have been created for digitized Islamic manuscripts, with their own portals. One draws from Princeton's own holdings, the other is thematic, focussed on manuscripts from Yemen. The former was largely done through internal funds. The latter was supported as a project by NEH and DFG grants. Together they offer more than 1,500 manuscripts, of which the bulk is digitized microfilms. The metadata is useful, but not curated or cleaned up, meaning that there are too many categories and in addition they are not listed alphabetically. This is solved by the search function, which is excellent. For example, searching for "20 lines" gives all manuscripts whose text is laid out on 20 lines per page. It is beneficial to search in both Arabic and Latin transliteration as in rare cases the title is only mentioned in one or the other way. The viewer is simple but efficient, similar to the viewer of Archive.org (see below, under McGill University). A button for downloading the entire document is missing. In the case of the Yemen repository, this is remediated by having a table of contents as a sidebar when examining the manuscript in the viewer. This is a rather unique feature that greatly expedites and enhances the study of these manuscripts.

2.4 *University of Michigan Library*

The most exemplary digital repository of North America was done at the University of Michigan. More than 1,000 manuscripts were cataloged and digitized in a concerted effort, thanks to an NEH grant. They decided to leave out a portal and simply let users interact with the catalog, which is integrated into

[3] On Mac a screenshot can be made with: Shift+Command+3, on Windows and Linux: PrintScreen.

their wider library catalog, and with the repository, which they arranged to be hosted with HathiTrust. The library catalog is connected with the repository, which is useful as the search function works well in the catalog but is broken within HathiTrust. The viewer has a fair number of options, but lacks clear indication of which manuscript and folio the user is seeing. Several download options are available. In the cataloging process they experimented with crowdsourcing, which did not yield a significant result.[4]

2.5 *Islamic Studies Library, McGill University*

The repository at McGill distinguishes itself by focussing on lithographs, including only few manuscripts. Additionally, the books published by the Tehran branch of the Institute of Islamic Studies at McGill University have also been digitized and are available through the same portal, bringing the total to over 400. The digitization seems to have been done over the course of several years. The portal is hosted at the Internet Archive, a non-profit organization that seeks to archive a variety of things in digital format and make them freely available online.[5] The most obvious benefit of this is that this repository has a greater assurance of a continued existence, with hardware and software upgrades along the way. The metadata seems to have been pulled directly from the library catalog. There is relatively rich metadata, yet much of it is non-uniform. For example, not all manuscripts are tagged as manuscripts and there is both an "Islamic philosophy" category as well as an "Islamic philosophy." category, the difference being merely the period at the end, as well as nine categories all variations on the word "Philosophy." This will surely leave a user missing relevant texts when looking through one of the categories. Next to finding texts through filtering categories, one can also search by key words. Unfortunately, this search function seems to be unreliable. For example, searching for كاشف الأسرار, *Kashif al-asrar* and *Kāshif al-asrār*, all result in zero results, whereas there is a book available with this title. Users are able to view online, manually download a page, and download a PDF of the entire object. For downloading manually, first zoom in and only then right click and save, since this gets you a higher resolution image.

[4] See Kropf, E., Rodgers, J., "Collaboration in Cataloguing: Islamic Manuscripts at Michigan," pp. 17–29 in *MELA Notes* 82 (2009).

[5] They state on their website that: "The Internet Archive is a 501(c)(3) non-profit that was founded to build an Internet library. Its purposes include offering permanent access for researchers, historians, scholars, people with disabilities, and the general public to historical collections that exist in digital format." "About the Internet Archive", archive.org.

3 Major Collections in the MENA Region

In the Middle East and North Africa we can find some of the largest digitization efforts. Especially notable are the efforts of libraries in Istanbul and Tehran to conserve their heritage; almost everything in the centralized collections of these two cities has been digitized or can be digitized upon request. To that we may add the efforts at King Saud University in Riyadh which has engaged in creating an online repository of unmatched dimensions, and the National Library of Morocco which is on its way to establish a sophisticated digital flank to their library system. Other cities with notable collections, such as Cairo, Damascus, Najaf, and Baghdad, are lagging behind. Notwithstanding, the efforts by the four aforementioned cities has been truly transformative for our conduct of research and they merit further attention.

3.1 *Süleymaniye Kütüphanesi*

Currently, the Suleymaniye Library sits on top of the largest pile of digitized Islamic manuscripts in the world, likely in the range of 100,000. Most collections in Istanbul were centralized in the Suleymaniye and from the early 2000's onwards they started to digitize basically everything. The size of the collection made this a long process, and this, in turn, causes significant quality variance among the corpus, as over the years different equipment and technologies were used to create digital surrogates. These photos are not free, nor available online, but can be accessed on computers at Suleymaniye itself, or can be send upon request. It seems that at time of writing they are busy to make the photos freely available online. The catalog is notorious for utilizing an inconsistent Turkified Latin transliteration. They are actively curating their digital catalog and repository, so these shortcomings will hopefully be taken care of in the future.

3.2 *Topkapı Sarayı Müzesi Kütüphanesi*

I mention Topkapi separately since they operate independent of Suleymaniye and enforce a very strict policy towards digitization. In 2014, I had to go in person and get written permission from the Ministry of Culture and Tourism to even file a request. It turned out my manuscript had not been digitized, but they were willing to do so, without an additional fee (the per photo fee can quickly add up though). A few weeks later I received the files. As such, I include this collection to show that sometimes digitized manuscripts cannot even be consulted on-site, but have to be obtained by exerting a lot of time, energy, and money, without even perusing an on-site catalog.

3.3 *Ketābkhāna va mūza-ye melli-ye Malek*

Malek Library, in Tehran, provides free, online access to digital surrogates of their manuscripts. To date, more than 5,000 manuscripts have been digitized, according to their website. The online catalog is among the most extensive, with ample metadata for virtually every holding. The portal is in Persian only, and allows for detailed searches on every individual element of the metadata. The site can be slow and can time-out at times. Not all manuscripts can be viewed, and for ones that can, the viewer is basic. You get to see one photo and need to use buttons to move to the next or previous page. This much is available for free, beyond that requires a log-in.

3.4 *al-Maktaba al-waṭaniyya li-l-mamlaka al-maghribiyya*

The National Library in Morocco currently has about 100 manuscripts available online. They, supposedly, set out to digitize all of their manuscripts, 80,000 in total, to be recorded in a digital catalog. Manuscripts are not their sole focus, as they include a variety of other library holdings such as journals and even audio recordings such as audio books. As such, this is a multi-year endeavor, which looks to be a pillar of the library's long-term strategy. Previously, the library had put large amounts of manuscripts on microfilm, requiring visitors to look at the microfilms rather than the original objects. As these objects are digitized, visitors are now allowed to see these digital surrogates.[6] With only a hundred online, the free dissemination seems to be not a priority. Notably, the viewer is built with Microsoft Silverlight, a technology similar to Adobe Flash, meant to deny the possibility of downloading the images. One cannot zoom in very far, and browsing can only be done page by page.

3.5 *Jāmiʿa Malik Saʿūd*

The King Saud University is in possession of some 11,000 manuscripts. So far, about half of them are made freely available on a dedicated website, with decent metadata which is also searchable. One does well to navigate the website in Arabic, as the English version seems less reliable. The portal, in general, is solid. The breadth of topics included is commendable, covering virtually all aspects of Islamic civilization. The viewer is too basic, offering no flexibility beyond looking at a page, clicking on it to see a larger version, and offering the option to see the page as a PDF. Furthermore, every image has a big watermark. On certain parts of the internet, files can be found of entire manuscripts of this

[6] Hendrickson, J., Adil, S., "A Guide to Arabic Manuscript Libraries in Morocco: Further Developments," pp. 1–19 in *MELA Notes* 86 (2013), p. 5.

collection, though the quality of those files may be less than currently found on the official website.

4 Notable Collections in Africa, the Levant, and Asia

Under this header I would like to shed light on the variety of digital repositories that exist beyond the beaten path of the famous collections described above. I chose to divide this up into three regions: Africa, the Levant, and Asia. South America, Oceania and Antarctica remain completely absent from this chapter, as they are largely without Islamic manuscripts. The three chosen regions, on the other hand, all have a strong history with Islamic culture and therefore have a large amount of manuscripts. Unfortunately, all three regions sorely lack in digital resources for these manuscripts.

For all three regions I specifically sought for local digital repositories; conceptualized and executed at the particular place of a collection. Much more work could and should be done on the collections from these regions. A collaborative approach may be unavoidable, such as done by the British Endangered Archives Project, the German Gerda Henkel Foundation, and the Hill Museum and Manuscript Library (see the Timbuktu repository). From the libraries selected in this article, it should become clear that the vast area in between Tehran and Tokyo is lagging behind in disseminating digitized manuscripts. My personal experience with Punjab University Library, in Lahore, Pakistan, speaks to this. It took me nearly half a year of e-mailing and eventually having someone go in person, and then paying a fee, to obtain photos of only mediocre quality. For India, it seems that the organization National Mission for Manuscripts has been industriously digitizing manuscripts. They do not, however, state how these digital surrogates can be obtained, and their website is offline at time of writing. The Internet Archive makes snapshots of websites and their last record is from September 2016, when the NMM's website claimed to have digitized close to 3.5 million manuscripts. We can only hope these files will be made public at some point. In the meantime, let us consider the following repositories.

4.1 *Jāmiʿa al-najāḥ al-waṭaniyya*

Admirably, the university in Nablus, Palestine, makes more than 700 of their manuscripts freely available online. Some basic errors will make it hard to profitably make use of it, though. Most notably, no mention is made of a call number, making it difficult to refer to them. A large part of the collection consists of digitized microfilms. The catalog includes only the title and author, and clicking on them immediately opens the viewer. Viewing can only be done one page at a time. Clicking on an image opens a much higher quality photo on which the watermark is relatively small.

4.2 *Jafet Library*

The digital repository at the American University of Beirut is nothing more than a page with titles and clickable covers. Clicking on them takes you to a page with no more information than the title and author—no MSS number is mentioned. The viewer is as simple as it can get, and only allows access to the first and last five pages of a manuscript. Since only 27 manuscripts are listed, it looks more like a pilot project than anything else.

4.3 *Aboubacar Bin Said Library and Mamma Haidara Library on vHMML*

The collections of Timbuktu made world news when it seemed to become the next target of cultural heritage destruction by rebels, in 2013. As is clear from the repository here under consideration, this did not happen, or at least not completely. SAVAMA-DCI,[7] a Timbuktu-based NGO, collaborated with several organizations to transfer the manuscripts to safety and digitize them. The Hill Museum and Manuscript Library, at Saint John's Abbey and University, has been the primary institution to take on the task of digitization and building a repository. Of the Timbuktu collections, about 1500 are available online, with the promise of possibly the entire collection of several hundred thousand manuscripts to be digitized. They are available through the portal called vHMML. The portal and viewer are modern and flexible from a technical point of view, but can be notably slow in use. Moreover, if you do not search for a specific title, titles are not shown in the list view. Instead, the user only sees the name of the collection, which is rather useless information. This makes casual browsing in hopes of a serendipitous encounter nearly impossible. On the same portal, other relevant collections are also available, most notably from Lebanon and Jerusalem.[8] Combined, the digital collection comes to about 5,500 manuscripts. When you combine manuscripts tagged as Arabic, Kurdish, Persian, or Turkish, this number rises to almost 8,500. In short, it seems that vHMML is fast becoming one of the primary destinations for those interested in Islamic manuscripts. Manuscripts can be viewed online, after free registration. Downloading is behind a paywall. Who the actual copyright owner is, and who will receive the money paid, remains unclear. This is, I think, a precarious

[7] The full name is L'organisation Non-Gouvernmentale pour la Sauvegarde et la Valorisation des Manuscrits pour la Defense de la Culture Islamique.

[8] HMML carefully avoids labelling Jerusalem as either Palestine or Israel.

issue. The West has had a painful history of claiming cultural heritage in foreign countries and shipping it off without compensation. Such issues should be thought out in the digital sphere too. I am not saying HMML is in the wrong here, but merely wish to note the lack of clarity in their documentation. It should also be noted that other organizations involved in the digitization of the Timbuktu manuscripts are more suspect in this regard. Aluka, a project of a US-based NGO, offers more than 300 manuscripts on Jstor, behind a paywall.[9] Another project, supported by the Gerda Henkel Stiftung and the University of Cape Town, supposedly hosts digital surrogates but restricts access to registered users, and has quietly turned off the registration form.

4.4 *MyManuskrip Malaya University*

I mentioned earlier that between Tehran and Tokyo, very few repositories are to be found. A fitting example for this claim is that the only free, online repository of Islamic manuscripts I could find is no longer available. The Malaya University in Malaysia used to run a website called MyManuskrip, on which they hosted a significant amount of manuscripts, in a variety of languages among which was Arabic. It seems to have been put reasonably well together, that is to say, with effort scholars could have made profitable use of it. The project was started with much enthusiasm in the late 2000s.[10] The plug was pulled somewhere in 2014 or '15, without warning or explanation. All that is left are archived snapshots of some pages of it. As such it is a sober warning that on the Internet, what is one day, can be not the next.

4.5 *Daiber Collection Database at the Institute of Oriental Culture, University of Tokyo*

Apparently the German scholar Hans Daiber was an avid collector of Islamic manuscripts. His collection ended up in Tokyo, and the Japanese did a good job providing digital access to more than 500 of them, likely the entire collection. The website which hosts it looks and feels barebones, but it makes up for this in versatility. For example, because of the way this website is put together, it is one of very few collections that allows users to search for specific characteristics, such as the number of lines of text a manuscript has. I also greatly appreciate the distinction between manuscript and text; they basically treat every manuscript as a collection of texts (*majmūʿa*) and subsequently list the

[9] Cf. Ryan, D., "Aluka: digitization from Maputo to Timbuktu," pp. 29–38 in *OCLC Systems & Services: International digital library perspectives* vol. 26, iss. 1 (2010).

[10] Zaynab, A.N., A. Abrizah, and M.R. Hilmi. "What a Digital Library of Malay Manuscripts Should Support: An Exploratory Needs Analysis." pp. 275–289 in *Libri* 59 (2009); Zulkifli, Z. "A Collaborative E-Workspace for Digital Library of Malay Manuscripts." pp. 368–372 in *International Journal of Information and Education Technology* 4, no. 4 (2014).

different texts contained within, even if only one text fills the entire manuscript. This conceptually separates the codicological description of the object and the philological description of its contents.

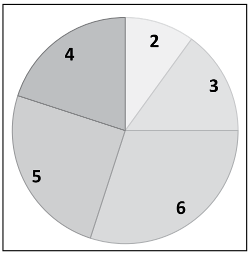
5 A Grand Comparison of the Quality of Digital Surrogates

Now that we have come to know various repositories, let us compare them, so that we may better understand the state of the art of digitization of manuscripts worldwide. I shall do this along the ten notions discussed in the previous chapter, after which I will make some final remarks on the general result of our analysis. The notions are: size of the collection; online availability; ability to download; the portal; the viewer; indication of page numbers; image resolution; color balance; lighting; and how the image is cut. For each of them I prepared a pie-chart which is easiest read starting top-right and going clockwise. The slices present the number of repositories which perform from bad to good under the specific category. In listing the repositories I have used shortened names which I hope are self-explanatory.

5.1 *Size of Collection*

Digitization is about scaling up. Digital photos can be stored and transmitted cheaply in large quantities. Maintenance, including migration and upgrading, can be done as easily for a large set of files as it is for a small set. Larger repositories can therefore benefit from a larger total pool of funds-per-file, ensuring a more professional and future-proof curation. In grading I looked in particular at how much has been digitized in comparison to the total collection of actual manuscripts held by the library.

TABLE 3.1 Size of collection



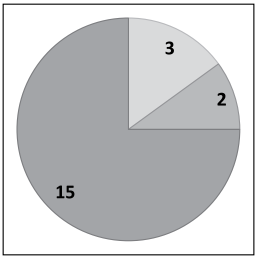
(Almost) nothing:	AUB, Malaya
Only little:	Vatican, Leiden, Morocco
Quite a few:	BL Qatar, Beinecke Yale, HMML Timbuktu, Princeton, UC Los Angeles, Topkapi Istanbul
A good amount:	BnF Paris, Stabi Berlin, McGill Montreal, Malek Tehran, Najah Nablus
A lot:	Michigan, Daiber Tokyo, King Saud Riyadh, Suleymaniye Istanbul

It is good to see the top two categories make up close to half of all the repositories. Of the second category, which only offer a little, we know that the Vatican and the National Library of Morocco intend to make much more digitally available. The national libraries of France and Germany are likewise still expanding their digital holdings. As such, the future is looking bright. Notably, of the middle category we have less expansion to expect, since most of them digitized on a project basis which has already terminated.

5.2 *Online*

Digitization inevitably means online distribution. Given that classical Islamic studies remains a niche discipline, free distribution is essential. Allowing re-distribution, for example by licensing the photos in the public domain, is much welcomed too.

TABLE 3.2 Online availability

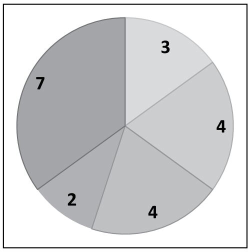
	Not online:	Topkapi Istanbul, Suleymaniye Istanbul, Malaya
	Behind paywall:	Leiden, Malek Tehran
	Free online:	BnF Paris, Stabi Berlin, BL Qatar, Beinecke Yale, Michigan, Daiber Tokyo, Vatican, Princeton, AUB, UC Los Angeles, McGill Montreal, King Saud Riyadh, Morocco, Najah Nablus, HMML Timbuktu

I divided the repositories into three groups. Those that are not online, meaning you have to order photos at the library itself, for example on a CD, or they do not exist anymore, in the case of Malaya University. Those for which you have to pay in order to instantly see them online. And those which are fully, freely accessible online. As becomes clear from the pie-chart, we live in a world in which most repositories make their Islamic manuscripts freely available.

5.3 *Downloadable*

Are the digital surrogates downloadable? This is different from the previous category. For example, Leiden's manuscripts are behind a paywall, but downloadable, while UCLA's manuscripts are freely accessible online but not downloadable. For the first category downloading is impossible since they are not

TABLE 3.3 Function for downloading



Impossible:	Topkapi Istanbul, Suleymaniye Istanbul, Malaya
Viewing only:	UC Los Angeles, Morocco, Malek Tehran, HMML Timbuktu
With watermark:	Daiber Tokyo, Vatican, King Saud Riyadh, Najah Nablus
Per photo:	Princeton, AUB
Full manuscript:	BnF Paris, Stabi Berlin, BL Qatar, Beinecke Yale, Michigan, McGill Montreal, Leiden

made directly available online. Libraries which allow online viewing and put in place technology to prohibit users from downloading are in the second category. One can still make a screenshot, of course, if all that is needed is a small part. Since this is too labor intensive to do for even a small, relevant portion of a manuscript, these manuscripts can be said to be undownloadable. Next are those for which a water mark is included in the downloaded photos. A higher category allows users to download one photo at a time. The highest category, then, means that users can download the entire manuscript, and often times there will be functionality to download only parts of it.

An important part of this aspect of being downloadable is the restrictions that libraries impose on using the digital surrogates, that is, their policy on copyright. In principle, anything on the web can be captured and independently distributed, and often times this is much easier done than the independent reproduction of paper copies of books and manuscripts. It is therefore good to avail ourselves beforehand of the intended use, to forego any legal problems. It turns out there is a vast difference in how libraries are situating their repositories of digital surrogates of ancient manuscripts. In total, I distinguish six attitudes on redistribution.

The first is to assert that the photos are in the public domain, thereby allowing any type of usage and redistribution. The repositories of the University of Michigan and the Qatar Digital Library do this.

A tier below that is allowing non-commercial use, as long as the source is attributed to the library. The Staatsbibliothek Berlin, Biblioteque nationale de France, and American University in Beirut assert this. Next is Tokyo, which only allows downloading and transformation for personal use.

Then there are those who under all circumstances require to be asked permission. These are the Vatican, the Leiden-Brill project, the Timbuktu manuscripts on vHMML, and the Suleymaniye and Topkapi libraries. Brill, it should be noted, gives some leeway to download and transform for personal and academic usage.

This last remark may sound vague, but that is nothing compared to the next category. This category comprises libraries who purposely describe the rights and permission in vague and opaque terms. Since I do not know what they mean to us on a practical level, I will not summarize it. It seems that their statements are not written for users, but rather to safeguard themselves from any litigation. These are the repositories of Princeton, the Beinecke at Yale, and McGill.

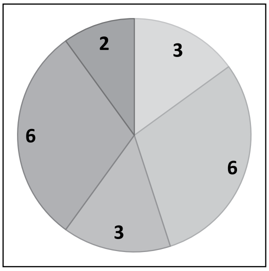
Lastly, the repositories of Malaya University, Najah in Nablus, King Saud in Riyadh, and UCLA do not give any terms of use. I find it interesting that of the last three categories, it is the commercial enterprise, Brill, which is the most friendly to reuse and gives the best options for contacting them. I therefore mean to exclude Brill when I say that using repositories of the last three categories is, on paper, a risky undertaking. Legal action against perceived misuse of academic materials is becoming more common. The question 'Will I be sued if I do something with these images?' deserves a simple answer. To have institutions with big pockets like Princeton and Beinecke answer with 'Maybe. Maybe not.' is discomforting to say the least. Taken as a whole, though, it seems well within virtually any country's legal code, and within most libraries' explicit policy, to keep private copies on personal computers for research purposes.[11]

5.4 *Portal*

Digitized manuscripts are usually accessed through what I call a portal and a viewer. A portal is a website featuring a catalog, which allows a user to find a particular manuscript. The viewer is a page which allows a user to view the photos of a manuscript, usually with buttons to navigate and zoom. The portal is judged on how well it discloses the digital holdings. Given its importance, I already gave it attention in the description of each repository.

[11] This is admittedly a grey area in which, it seems, the code of law itself is inadequately covering these issues, cf. Besek., J.M., et al., "Digital Preservation and Copyright: An International Study," pp. 104–111 in *The International Journal of Digital Curation*, vol. 2, nr. 3 (2008).

TABLE 3.4 Usefulness of the portal



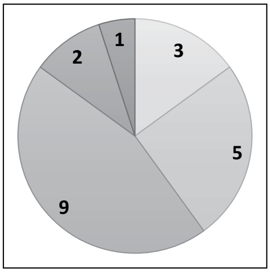
Missing:	Topkapi Istanbul, Suleymaniye Istanbul, Malaya
Poor:	Michigan, Vatican, AUB, UC Los Angeles, Leiden, Najah Nablus
Satisfactory:	Beinecke Yale, Morocco, Malek Tehran
Good:	Stabi Berlin, BL Qatar, HMML Timbuktu, Princeton, McGill Montreal, King Saud Riyadh
Excellent:	BnF Paris, Daiber Tokyo

In general, the portal seems to be an afterthought. An extreme example of this is the case of the Michigan University repository. They first had a dedicated portal and later deleted it, obliging users to find their own way within the larger online library catalog. On the other end of the spectrum we may note portals such as for the Daiber Collection Database and the Qatar Digital Library, which neatly define their purpose and bring all relevant information together. As I noted before, in the latter case there is even value added by including research articles within the portal to showcase its usability.

5.5 *Viewer*

The viewer is judged on its flexibility in viewing and navigating the manuscript. It too was highlighted above in the description of the repositories.

TABLE 3.5 Usefulness of the viewer



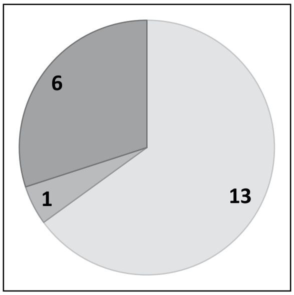
Missing:	Topkapi Istanbul, Suleymaniye Istanbul, Malaya
Poor:	Daiber Tokyo, AUB, Morocco, Malek Tehran, Najah Nablus
Satisfactory:	Beinecke Yale, Michigan, HMML Timbuktu, Vatican, Princeton, UC Los Angeles, McGill Montreal, King Saud Riyadh, Leiden
Good:	Stabi Berlin, BL Qatar
Excellent:	BnF Paris

A satisfactory grade was given to indicate that a user can generally make use of a repository but will experience inconvenience and will like have to put in extra effort. Beggars can't be choosers so I gave it a mildly positive term, but in reality this is not a good grade. Using technology only makes sense insofar as it *helps* us. When it starts working against us, which is the case for the vast majority, we can only proceed with caution. Given this result, the viewer is in my estimation a point on which all repositories can greatly improve. They may take the Gallica Viewer of the Bibliothèque nationale de France as an example.

5.6 *Page Numbers*

Is the user able to quickly deduce which folio and which manuscript they are seeing? It turns out, many repositories forget to build in this functionality, while it is crucial for looking something up or citing it. When pencilled folio numbers are absent from the material manuscript, it is arduous and prone to error to count back from the first folio in a digital surrogate. Many repositories give false information, by providing a 'page number' which is actually an image number, starting from the very first image which is usually the cover or an index card of the catalog.

TABLE 3.6 Indication of page numbers



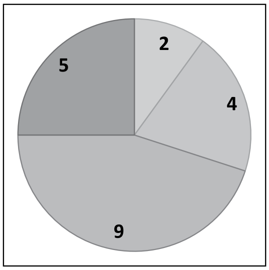
No:	Michigan, HMML Timbuktu, Daiber Tokyo, Princeton, McGill Montreal, King Saud Riyadh, Leiden, Morocco, Malek Tehran, Najah Nablus, Topkapi Istanbul, Suleymaniye Istanbul, Malaya
Some:	BnF Paris
Yes:	Stabi Berlin, BL Qatar, Beinecke Yale, Vatican, AUB, UC Los Angeles

Please note that I drew up these answers by taking sample tests for each repository. I discovered in the case of the Bibliothèque nationale de France that page numbering can be correct for certain items but missing or incorrect for others. This may likewise be true for other repositories. I stand by my conclusion, though, that page numbering is among the biggest problems repositories currently have. Notably, none of the repositories include a feature which I think would be highly useful, which is to automatically generated a small stamp in a corner of the image indicating its origin. Such information would ideally consist of library name, manuscript number, and folio number. If this would be done, the origin would be hardcoded into the image, making it easier for users to use individual photos or reorganize them on their computer. A similar system could be put in place for the file name when downloading images, which is again an opportunity missed by all repositories. Once origin details are robustly taken care of, scholars will find it much easier to use and refer to digital surrogates in their writings.

5.7 *Resolution*

Under this header I judge the level of detail of digital surrogates, based largely on the file size in combination with the dimensions of the picture (height and width in pixels). Both should be taken into account because there is invariably some form of compression, and hence loss of quality, involved in the process from making photos to putting them online. I have come to think of 500kb per page as a desired minimum, meaning 1mb for a two-page spread. A visual assessment is also informative, as different file formats make it hard to objectively compare all repositories. Later in this chapter we shall take a look at a visual comparison to give an impression of the quality.

TABLE 3.7 Image resolution

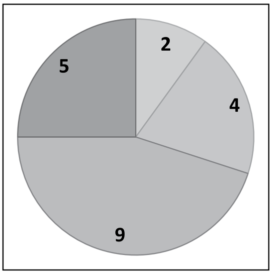
	Poor:	Malek Tehran, Malaya
	Satisfactory:	King Saud Riyadh, Leiden, Morocco, Najah Nablus
	Good:	Beinecke Yale, Michigan, Daiber Tokyo, Vatican, Princeton, AUB, UC Los Angeles, McGill Montreal, Suleymaniye Istanbul
	Excellent:	BnF Paris, Stabi Berlin, BL Qatar, Topkapi Istanbul, HMML Timbuktu

In my opinion, the vast majority of digital surrogates are fine to be used for various scholarly purposes. Interestingly, this division seems to correspond to a geographical division. In the top category comes Europe. In the second category North-America, in the lower categories the rest of the world. Notable exceptions are Topkapi and Leiden. The former does an excellent job, as far as I have been able to assess. The latter is lagging behind, which is all the more surprising as it is the only for-profit repository.

5.8 *Color Balance*

In the graphic design industry, people sometimes ask if the blacks are black. They mean that if an object is in reality black, does it display on a computer screen as black or does it seem like dark grey? True-to-color is also an issue for

TABLE 3.8 Color balance



Poor:	King Saud Riyadh, Suleymaniye Istanbul
Satisfactory:	UC Los Angeles, Malek Tehran, Najah Nablus, Malaya
Good:	BL Qatar, Daiber Tokyo, Vatican, Princeton, AUB, McGill Montreal, Leiden, Morocco, Topkapi Istanbul
Excellent:	BnF Paris, Stabi Berlin, Beinecke Yale, Michigan, HMML Timbuktu

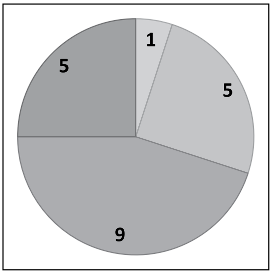
us, as the color of the paper, ink, and binding are part of the materiality of the manuscript and we do well to preserve them truthfully in their digital surrogate. This is what I assess under this heading.

Most repositories do well with this. The difference between the excellent and good categories was made on consistency; the 'good' repositories show their manuscripts true to color but have an occasional, slight aberration, whereas the 'excellent' repositories are consistent. At this moment, color balance does not need to be a concern for us, for most purposes. One case when we should keep it in mind is when we use digital surrogates from different repositories that have a different color balance quality.

5.9 *Lighting*

Ideally, photos are made in a well-lit place so that the entire folio is evenly visible. When one, direct light source is used, it often creates shadows and shines, which can make it difficult to read parts of the text and give an uneven, nervous quality when looking at the entire page.

TABLE 3.9 Lighting



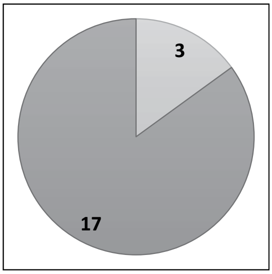
Poor:	Najah Nablus
Satisfactory:	McGill Montreal, King Saud Riyadh, Malek Tehran, Suleymaniye Istanbul, Malaya
Good:	Michigan, Daiber Tokyo, Vatican, Princeton, AUB, UC Los Angeles, Leiden, Morocco, Topkapi Istanbul
Excellent:	BnF Paris, Stabi Berlin, BL Qatar, Beinecke Yale, HMML Timbuktu

As with color balance, the majority of repositories does well for lighting. However, in this case the bottom two categories are concerning. For these, six repositories in total, shadows and shines pose a real threat to the readability (and hence usability) of the manuscripts.

5.10 *Cut*

The borders of a photo crucially separate what remains digitally visible, and what is out of sight. I have given a high score if at all times the entire edge of the manuscript is visible, in other words, if also a bit of the supporting surface can be seen. Lower scores are given for when a photo is cut tighter. Sometimes they are too tight, when marginalia are cut off.

TABLE 3.10 Cut

	Poor:	McGill Montreal, Najah Nablus, Malaya
	Excellent:	BnF Paris, Stabi Berlin, BL Qatar, Beinecke Yale, Michigan,HMML Timbuktu, Daiber Tokyo, Vatican, Princeton, AUB, UC Los Angeles, King Saud Riyadh, Leiden, Morocco, Malek Tehran, Topkapi Istanbul, Suleymaniye Istanbul

At the start of my investigation, the cut was a major concern to me. It is a relief to see that virtually all repositories get this right. In some cases, such as many manuscripts from Suleymaniye, the cut is very tight, but not too tight.

5.11 *A Final Rating*

We can combine the previous evaluations in one grade, to get a sense of the overall performance. For four categories, they were assigned values from 0 to 4. In case of three categories they received the values 0, 2, or 4, and where there were only two categories they were assigned 0 or 4. This means that repositories falling in a lower category are penalized considerably and this should be taken into account when evaluating the final grade. The scores were transformed to a grade on a 10-point scale and to a letter grade. We get the following list:

TABLE 3.11 Ranking of twenty repositories

Position	Library	Final grade	Letter
1	**Bibliothèque nationale de France**	9,3	A
1	**Staatsbibliothek zu Berlin**	9,3	A
3	**Qatar Digital Library**	8,8	B+
4	**Beinecke Rare Book & Manuscript Library**	8,3	B
5	**University of Michigan Library**	7,3	C
6	**Timbuktu Libraries on vHMML**	7,3	C
6	**Daiber Collection Database at the Institute of Oriental Culture, University of Tokyo**	7	C-
8	**Biblioteca Apostolica Vaticana**	6,8	D+
8	**Princeton University Library**	6,8	D+
10	**Jafet Library**	6,5	D
10	**Caro Minasian Collection at the University of California, Los Angeles Library**	6,5	D
12	**Islamic Studies Library, McGill University**	6,3	D
13	**Jāmiʿa Malik Saʿūd**	6	D-
14	**Leiden Universiteitsbibliotheek**	5,5	F
15	**al-Maktaba al-waṭaniyya li-l-mamlaka al-maghribiyya**	5,3	F
16	**Ketābkhāna va mūza-ye melli-ye Malek**	4,5	F
17	**Jāmiʿa al-najāḥ al-waṭaniyya**	4,3	F
18	**Topkapı Sarayı Müzesi Kütüphanesi**	4	F
19	**Süleymaniye Kütüphanesi**	3,5	F
20	**MyManuskrip Malaya University**	1,5	F

What is immediately visible is that the letter grades make less sense in this case. This is especially so since in letter grade education systems grades are usually arrived at by starting at A and then deducting when necessary, whereas in this evaluation the digital materials have to slowly earn a higher grade by virtue of doing better in a certain regard. Another important note is to bear in mind that a low grade does not mean the repository is unusable. It only indicates that it is lacking certain qualities that were included in this evaluation. Comparatively, however, it is highly instructive to notice how far ahead the Bibliothèque nationale de France and the Staatsbibliothek zu Berlin are, against the rest. They perform well both in terms of delivery (online availability, portal, viewer, etc.) and in terms of quality (resolution, lighting, etc.). This is true too of the Qatar Digital Library. The other two in the top five, Beinecke and Michigan, earn their place more on their quality than their delivery. Beyond the top five, we find a large group sitting between the grades 7 and 6. Leiden's repository is perhaps the most surprising. It is well beyond the top ten, and barely receives a passing grade. In the bottom part of the list, we may notice that Suleymaniye and Topkapi are punished for the difficulty of accessing their digital surrogates.

6 A Visual Comparison

To support the judgments I made in evaluating the repositories on the quality of their images, I include here a sample that illustrates the difference one can encounter while looking for photos of manuscripts. To make this as objective as possible, I set out to find in each collection a manuscript of about 20 by 15 cm, with about 19–23 lines of text per page, written in a straight *naskh*. From those manuscripts I selected an image that included the word *qāla* ('he said')

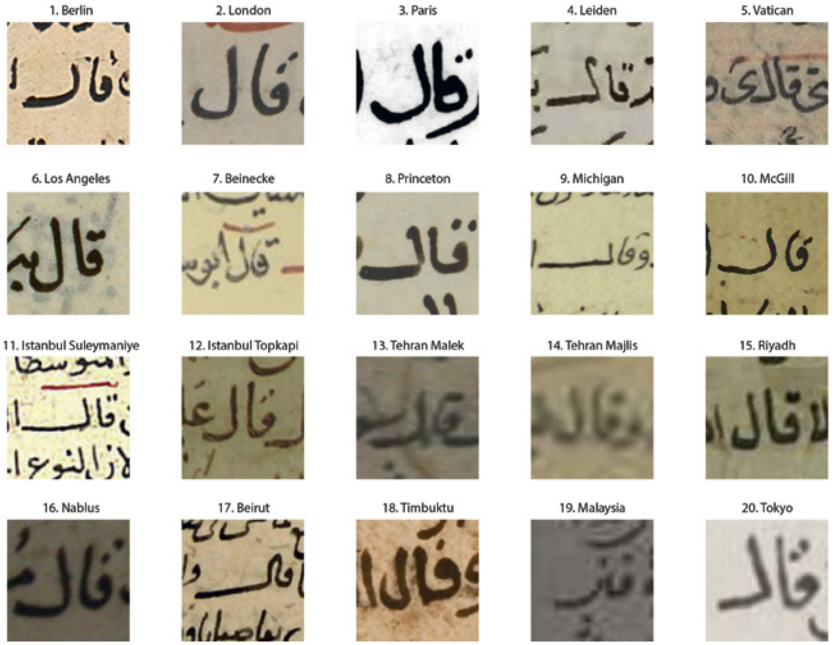


FIGURE 3.1 Comparison of *qāla* from different online repositories around the world

which I cut out in a square, and resized each square to be of the same size. I further did my best to obtain the highest quality color image in each case. For some repositories, I had to make concessions. For example, the American University in Beirut has so few manuscripts that I had to settle for a slightly less straight *naskh* without diacritics (*iʿjām*). In the case of Timbuktu, all manuscripts were copied in *Maghribī* script, being slightly more round and using only one dot to indicate a *qāf*. Further, the repositories of the BnF, Princeton, Nablus, Tokyo, and the Vatican use a lot of microfilms, resulting in black and white images, yet I chose a color image for the comparison. It should also be noted that for the Suleymaniye it is hard to let one image be a representative for the entire collection since there is considerable difference in quality.

7 Difference between Professional and Amateur Photos

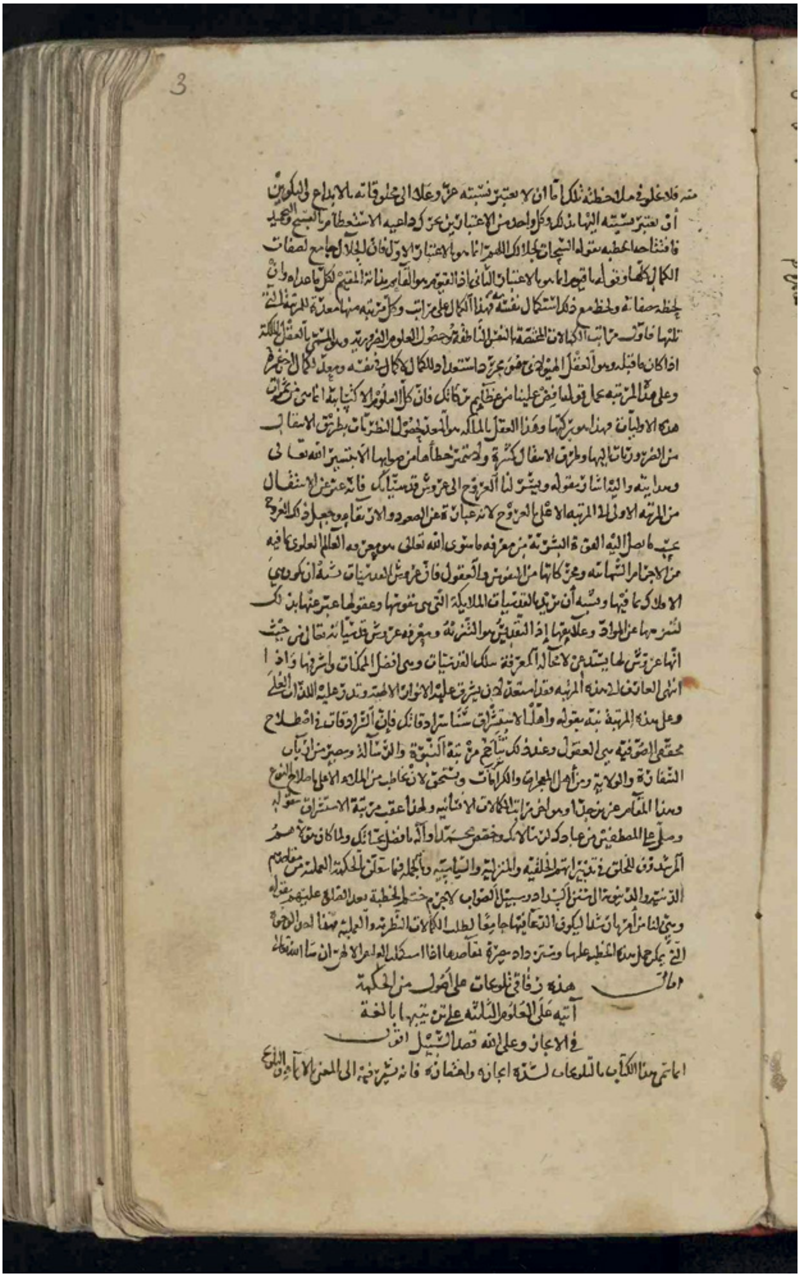
Relatively simple consumer electronics can also be used to make photos. On the next page we see on the left a professional photo of MS Leiden Or. 137, f. 3a (Ibn Kammūna, *Sharḥ al-Talwīḥāt*), and on the right a photo I shot myself, using an iPad.[12] At that time, I was not concerned to make precise shots for use in serious editorial work as I merely wanted to document the various manuscripts at Leiden that were of interest to me. The shots I made show the pages at an angle and with heavy shading. Additionally, I did not get all the edges of the page.

The professional photo is much better in many regards. However, when we look at the readability of the two photos, I cannot detect that much difference.

8 The Future of Digital Manuscript Repositories

Digitized manuscripts come in a great variety yet at the same time their digital existence also show some common traits. The opportunities these traits offer can be described through a SWOT analysis—a tool from economics to measure the strengths, weaknesses, opportunities, and threats of a business. Strengths and weaknesses are positive and negative aspects inherent to a business, opportunities and threats are positive and negative aspects of a business brought about by the environment it is in. In this case I consider the business to be the digitization of Islamic manuscripts.

[12] I used a 4th generation model, from late 2012, which has a 5MP camera.



3

منه فلا علو في ملاحظة ذلك إلا أن لا تعتبر نسبته عز وعلا إلى مخلوقاته بالإبداع والتكوين
أو تعتبر نسبته إليها منذ كل واحد من الاعتبارين عن كداعيه الاستعظام بالتسبيح والتحميد
فافتتاح الخطبة بقوله سبحان جلالك اللهم إنما هو بالاعتبار الأول لأن الجلال الجامع لصفات
الكمال كلها وقوله ما أقهر إنما هو بالاعتبار الثاني إذ القيوم هو القائم بذاته المقيم لكل ما عداه وإن
لحظه صفاته ولحظ مع ذلك استكمال نفسه ثم هذا الكمال على مراتب وكل مرتبة منها معدة للمرتبة التي
تليها فأول مراتب الكمالات المختصة بالنفس الناطقة حصول العلوم الضرورية وهو المسمى بالعقل بالملكة
إذا كان ما قبله هو العقل الهيولاني محض مجرد استعداد للكمال لا كمال في نفسه وبعد كمال الآخر هو
وعلى هذه المرتبة يحمل قوله لما فرغ علينا من عظائم منائك فإن كل العلوم إلا كسبية إنما هي من نتائج
هذه الأوليات فهذه الصورة كلها وهذا العقل بالملكة هو المعد لحصول النظريات بطريق الاكتساب
من الضروريات بتاليفها وطرق الاكتساب كثيرة والصعب خطأها من صوابها إلا بتيسير الله تعالى
وهدايته وإلى ذلك أشار بقوله ويسر لنا العروج إلى عرش قدسك فإنه عبر عن الانتقال
من المرتبة الأولى إلى المرتبة الأعلى بالعروج لأنه عبارة عن الصعود والارتقاء وجعل ذلك العروج
غير ما يصل إليه القوة البشرية من معرفة ما سوى الله تعالى هو عرش وهو العالم العلوي بما فيه
من الأجرام السماوية ومن كانها من النفوس والعقول فإن عرش القدسيات ينبغي أن يكون هي
الأفلاك بما فيها وشبه أن يريد بالقدسيات الملائكة التي هي نفوسها وعقولها عبر عنها بذلك
لتنزهها عن المواد وعلائقها إذ التقديس هو التنزيه ومعرفة عرش قدسيته تعالى من حيث
أنها عرش لها يستدعي لا محالة المعرفة بتلك القدسيات وهي أفضل الممكنات وأشرفها وإذا
انتهى العارف إلى هذه المرتبة فقد استعد لأن يشرق عليه الأنوار الإلهية ويترقى إلى اللذات العلى
وعلى هذه المرتبة نبه بقوله وأهلنا لاستشراق سنا سرادقاتك فإن السرادقات في اصطلاح
محققي الصوفية هي العقول وعند ذلك تتأخر مرتبة النبوة والرسالة ويصير من أرباب
الشفافة والولاية ومن أهل المعجزات والكرامات ويستحق لأن يخاطب من الملأ الأعلى بإصلاح النوع
وهذا المقام عزيز جدا وهو آخر مراتب الكمالات الإنسانية ولهذا عقب مرتبة الاستشراق بقوله
وصل على المصطفين من عبادك لرسالاتك وخصص محمدا وآله بأفضل تحياتك ولما كان هؤلاء هم
المرشدون للخلق في تدبير أمورهم الخلقية والمنزلية والسياسية وبالجملة فيما يتعلق بالحكمة العملية من مصالحهم
الدنيوية والدينية المشتمل كذا وسبيل الصواب لا جرم ختم الخطبة بعد الصلاة عليهم بقوله
وبين لنا من أمرها ما يكون الدعاء فيها جامعا لطلب الكمالات النظرية والعملية معا لدى الرحمة
التي يمكن حصولها لنا المطلع عليها وسترداده بجزئه مقاصدها إنما استكمل العلم إلا الحمد إن شاء الله تعالى
امالى هذه رقائق لوامع على الأصول من الحكمة
آتية على العلوم الثلاثة على ترتيبها بالغة
في الإيجاز وعلى الله قصد السبيل انتهى
إنما سمي هذا الكتاب باللوامع لشدة إيجازه واختصاره فإنه يشير فيه إلى المعنى إلا ما في البلوغ

FIGURE 3.2A Professional photo

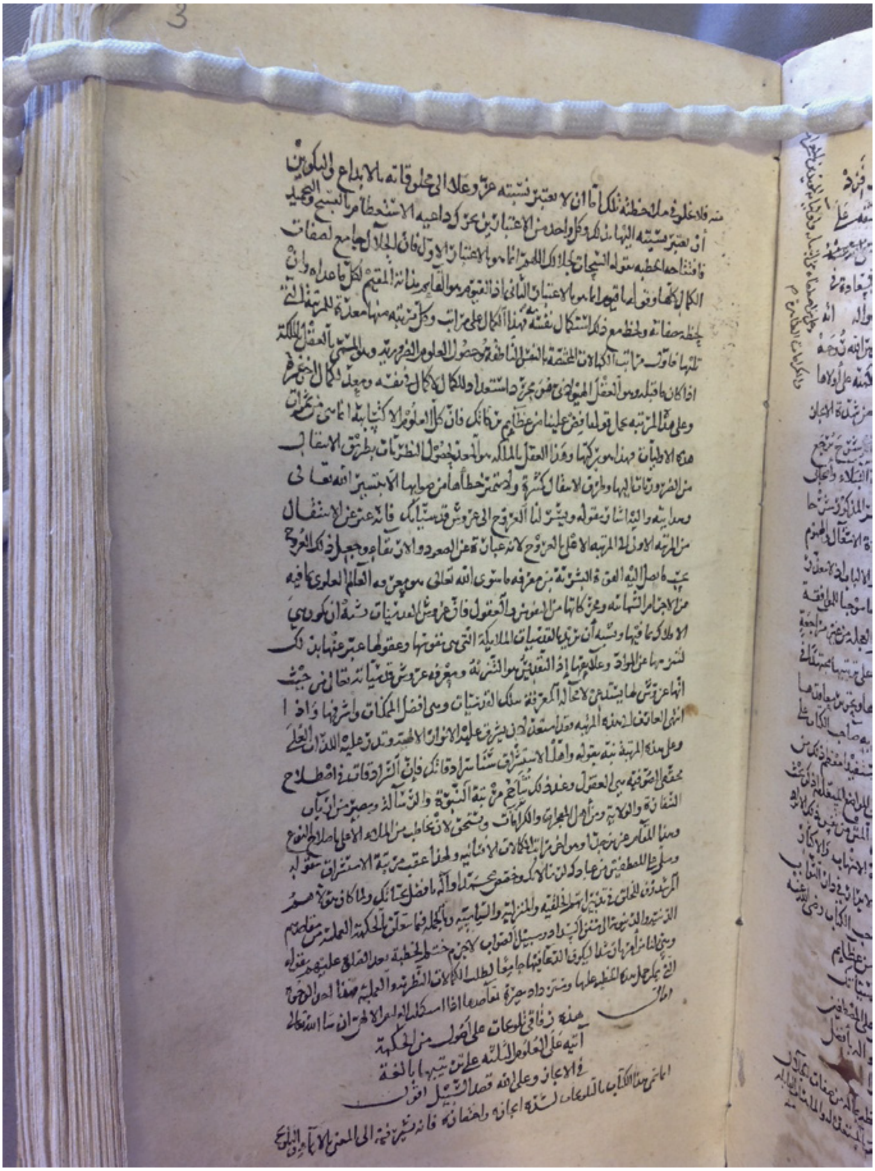


FIGURE 3.2B Amateur photo

Among the **strengths** of digitization of Islamic manuscripts worldwide is the big amount already done. Part of this, perhaps, is because of the presence of large amounts of Islamic manuscripts in countries with loose copyright laws and the availability of cheap labor. Although the quality of these photos varies tremendously, as we have seen, the vast majority is usable. I wish to draw particular attention to the study of *majmūʿa*'s.[13] Florilegiums are a particularly strong subset of Islamic manuscripts and they have largely been under-cataloged.[14] With their digital surrogates easier accessed, we can peruse, annotate, catalog, and study them much faster. What discoveries will come out of this, and the effect it will have on our methodology and general understanding of Islamic culture, remains to be seen, but I think it is among the most potent possibilities for a breakthrough, enabled by the digitization of Islamic manuscripts. A more obvious breakthrough, already gained, is that editions can now be based much easier on manuscript witnesses from all corners of the world and it is fairly easy to check existing editions against actual manuscript evidence. I do think, however, that editions based on digitized manuscripts require a description of the quality of the digital surrogates as part of the codicological description.

Weaknesses of digitized Islamic manuscripts pertain, I believe, to the seeming carelessness about securing that these digital surrogates are future-proof. I see two aspects to this. Firstly, the quality of some digital photos is lacking, sometimes spurred by a temptation to digitize microfilms. This results in poor surrogates, in fact, when a microfilm is used, we end up with a surrogate of a surrogate. The result may still be usable right now, but we do not know what we will want to do with them in the future. For example, technology is developed to automatically recognize text within manuscripts. For this to work properly, the quality of the photos is of paramount importance. Imagine if we have good optical character recognition technology, and are able to run this automatically over a large (huge, even) repository. We could engage in entirely new ways with the vast sea of literature that authors from the Islamic civilization produced! Since money is limited, and the number of manuscripts to be digitized is great, we realistically only have one opportunity to digitize a

[13] Nir Shafir also makes mention of this. Gratien, C., M. Polczyński, and N. Shafir. "Digital Frontiers of Ottoman Studies." pp. 37–51 in *Journal of the Ottoman and Turkish Studies Association* 1, no. 1–2 (2014), pp. 40–41.

[14] Some notable exceptions deserve to be mentioned, such as the catalogue dedicated specifically to the contents of florilegiums in Cairo's Dar al-kutub; Halwaji, A.S. (ed.), *Fihris al-makhṭūṭāt al-ʿarabīya bi-dār al-kutub al-miṣrīya: al-majāmīʿ*, London: Muʾassasat al-furqān li-l-turāth al-islāmī, 2011.

manuscript. By doing it poorly, we will likely hurt ourselves later on.[15] A similar weakness inherent to Islamic manuscripts is the relatively poor state of digital availability of catalogs. What good is a digitized manuscript if we cannot find it?[16] Further, having good, regularized metadata will open up possibilities of large scale research. Yet, we noticed that some repositories currently offer an overlapping, incoherent categorization and others describe their holdings in an eclectic, unpredictable transliteration. There is a solution to this, namely digitizing and regularizing printed catalogs, and this is a necessary condition to move the digital work on manuscripts forward. How much of this work ought to be the responsibility of scholars is not clear to me.

Secondly, no matter the quality, most digitization is done on a project basis, not as a program. For long term viability, the latter would be more desirable, seeing that there is continuity in funding, curation, and migration to newer technology.[17] When a project is finished, could it be merely waiting to go offline, like the repository at Malaya University, or become outdated and abandoned, like UCLA's repository? As with the quality of the photos, storage and curation needs to be thought of in terms of decades, for otherwise we are simply wasting our precious resources.[18]

What are some of the **opportunities**? With IT infrastructure and Internet connections becoming cheaper, the possibility is just around the corner for countries like Turkey, Iran, and India, to deposit their digitized manuscripts online, for free. This could make Islamic studies the field with the biggest amount of digitized manuscripts, in comparison to similar fields such as medieval studies, classics, and sinology. Already now, Islamic studies vies with other fields in the renaissance of philology within the humanities generally, as seen for example in the journal *Philological Encounters*. If in addition we can have more manuscripts digitally cataloged, perhaps connected through a union catalog, other possibilities will open up as well. For example, so far it has been nearly impossible to edit a text as a 'digital documentary edition,' in which the

[15] I agree with Jeanneney in that "haste does a disservice to the initiative." Jeanneney, J.-N. *Google and the Myth of Universal Knowledge: A View from Europe*. Translated by T.L. Fagan. Chicago: The University of Chicago Press, 2007, p. 55.

[16] As Weiss states in his analysis of massive digital libraries: "Without metadata available to anchor a digital version to the original object, it could [...] be 'lost at sea.'" Weiss, A. *Using Massive Digital Libraries*. Chicago: ALA TechSource, 2014, p. 28.

[17] Rieger, O.Y., "Enduring Access to Special Collections: Challenges and Opportunities for Large- Scale Digitization Initiatives," pp. 11–22 in *RBM* vol. 11, iss. 1 (2010).

[18] This point is also made in an excellent case study of digitization of African collections, see Krätli, G., "Between Quandary and Squander: A Brief and Biased Inquiry into the Preservation of West African Arabic Manuscripts: The State of the Discipline," pp. 399–431 in *Book History* 19 (2016), in particular p. 419.

reader can turn on and off manuscript witnesses to create particular readings that work for them, instead of relying on the one and only reading an editor offers through a traditional Lachmannian approach.[19] A documentary edition requires a broad set of manuscripts with strong metadata concerning their origin. As the methodology is being fine-tuned in other fields, we will soon be able to profit from this approach.

Secondly, if we find a stable way to store these digital surrogates, we could refer to manuscripts more easily in our studies. This is because this evidence would have a status of being semi-published, whereas before we could not expect our readers to have access to the manuscripts in order to fact-check our argumentation. This will require, again, more robust digital cataloging. One aspect much discussed in other fields is 'interoperability', meaning, agreements and technical frameworks that allow for easy communication between different repositories. The *International Image Interoperability Framework* is currently the most promising initiative in this regard, as discussed in Chapter Five. Through this framework, one is able to pull up and manipulate images of manuscripts of different libraries, side by side, in real time. It will be a long shot for libraries in the MENA region to join this initiative, but perhaps something similar will be developed, which is hopefully flexible enough to not have to reach manuscripts of each library in a specific way. Additionally, one could hope that this will encourage mirror-hosting agreements between institutions. A fine example of this principle is *The Stanford Encyclopedia of Philosophy*. This online encyclopedia is developed and hosted at Stanford University, but has agreements with the universities of Sydney, Australia and Amsterdam, the Netherlands, to run identical copies. This diversification strategy, literally spreading the odds over different continents, would be much welcomed in the world of digitized Islamic manuscripts.

Additionally, there is an opportunity for ameliorating digital surrogates by connecting them to large text corpora. Thus can be a two-way street. We could make connections to manuscripts from existing texts, and we could make connections from manuscripts to new entities within the text corpus. On a small scale, we are talking about a useful, flexible editing environment, but it could be done on a large scale in which case scholars could travel between texts by means of the digital text corpus and only dip into the manuscript evidence where necessary.

[19] For a critique on Lachmann's approach for Islamic texts, see Witkam, J.J., "Establishing the Stemma: Fact or Fiction?", pp. 88–101 in *Manuscripts of the Middle East* 3 (1988).

Lastly, let us consider the **threats**. I imagine these to be in the domain of continuity. As technology changes, our repositories are threatened to be left behind, eventually becoming inaccessible. Additionally, political pressure should be taken into account. Islamic studies is a field with clear political relevance in today's geopolitical discourse. As repositories seem to rely more often than not on state support, the state can weigh in on the course digitization takes. For example, a fair number of digitization projects in the US relied on grants from the National Endowment for the Humanities, an agency whose budget was threatened to be eliminated in '17– '18. It could also be that more power is asserted over repositories. A worrying first sign is that digitized manuscripts requested from Istanbul now come with a watermark, a new practice since 2017.

Whatever happens, it does seem that digitization has become an unstoppable force and more and more students and scholars are seeking out digitized manuscripts, often preferring them over material manuscripts. At the same time libraries are noticeably making it more difficult or even impossible to see the material manuscript, arguing that the digital surrogate will do. A fair evaluation of the digital materiality of digitized manuscripts is therefore crucial, in order to make use of them in an appropriate manner.

In this and the previous chapters, we have sought a better understanding of digitized manuscripts, which are digital documents with a strong relationship with material manuscripts that interact in a complex way with print publications. With this conceptual framework in place, it is now time to see what the essential skills and tools are when using digitized manuscripts. In the next chapter, we begin with this by considering the choice between working in a team or working alone, and by learning how to redraw glyphs and symbols in a more natively digital format.

CHAPTER 4

Paleography: Between Erudition and Computation

In this chapter, we shall go over three topics. First, we will look at how computer-supported research has developed within paleography, finishing with a close look at automated handwriting recognition. Then, we move on to discuss the drawbacks of costly team projects, which are all too frequent with research in digital humanities. Lastly, we shall look at how we can start doing paleography within a digital workflow by using ordinary consumer electronics and applications.

1 The Variety of Digital Paleographic Experience

Since the very beginning of modern paleography, with Jean Mabillon in the 17th century, there has been a tension between objective and subjective arguments, between evidence and erudition, between emphasizing systemization or diversity. A great example of such systematized collection of evidence is Adriano Cappelli's *Lexicon Abbreviaturarum*, first published in 1899, which remains until today the standard work for looking up abbreviations and their meaning for medieval Latin manuscripts. In 1978, Bernhard Bischoff went on record saying that "Mit technischen Mitteln is die Paläographie, die eine Kunst des Sehens und der Einfühlung ist, auf dem Wege, eine Kunst des Messens zu werden."[1] Today, I would argue that paleography is perhaps the most digital field of all of manuscript studies, even though the perception within the field is at times to the contrary.[2]

[1] Bischoff, B. *Paläographie Des Römischen Altertums Und Des Abendländischen Mittelalters*. Berlin: Schmidt, 1979, p. 17.

[2] See Stokes, P. "Computer-Aided Palaeography, Present and Future." pp. 309–337 in *Kodikologie und Paläographie im digitalen Zeitalter*, edited by M. Rehbein, P. Sahle, and T. Schaßan. Norderstedt: BoD, 2009, p. 321; Stansbury, M. "The Computer and the Classification of Script." pp. 237–249 in *Kodikologie und Paläographie im digitalen Zeitalter*, edited by M. Rehbein, P. Sahle, and T. Schaßan. Norderstedt: BoD, 2009, p. 238; Correa, A.C. "Palaeography, Computer-Aided Palaeography and Digital Palaeography: Digital Tools Applied to the Study of Visigothic Script." pp. 247–72 in *Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches*, edited by T. Andrews and C. Macé. Turnhout: Brepols, 2014, p. 247.

© CORNELIS VAN LIT, 2020 | DOI:10.1163/9789004400351_006

Computer-supported paleography finds expression in several ways. The main split is between those solutions which pertain to the publication phase and aim to publish or visualize something better, and those which pertain to the research phase and are meant to find data to answer a question.[3] Within the publication phase, there are roughly three different areas to be distinguished. First are those solutions that enhance conventional paper publications. Second are born-digital publications. Last, we find computer-supported innovation in education. As for the research phase, there are again three areas in which computers can enhance the work of paleographers. One is that a digital solution can be created, allowing a user to apply their traditional methodology more easily. Another is that a computer can execute the methodology itself and merely return back the final answer. Last, the digitality of digitized manuscripts and the computing power available can open up possibilities to come up with entirely new questions and new methodologies. All of these different ways of digital paleography depend on digitized manuscripts.

1.1 *From Paper to Digital*

The interaction between traditional publishing in paper form and the digital world goes both ways. Effort has gone into digitizing print materials, and techniques have been developed to digitally enhance a print publication.

For the former, a prime example is the previously mentioned lexicon by Cappelli. It was first published in 1899, after which a German edition came out in 1902, with a final edition released in 1929. There are several reprints until the digital world took hold. The original was scanned by the Google Books project and can be found twice on the net, seemingly since 2015.[4] The second German edition was scanned in Cologne and uploaded in 2012, with a copy uploaded to Archive.org in 2013. As a certain Mr. Degoix remarks on the weblog *The Ancient World Online*, the digital file is defective in several ways, mostly because it is simply missing certain pages. Another version of the second German edition, this time from 1901, was uploaded in 2015, again to Archive.org. Last, the English translation of the extensive introduction is also available online, hosted by the University of Kansas.

Remarkable it is, then, that the final edition of 1929 cannot be found. This did not deter further work on making this resource more digital in nature.

[3] Hassner et al. propose a different taxonomy, cf. Hassner, T., M. Rehbein, P. Stokes, and L. Wolf, eds. "Computation and Palaeography: Potential and Limits." pp. 1–30 in *Kodikologie Und Paläographie Im Digitalen Zeitalter 3*. Norderstedt: BoD, 2015, pp. 10–12.

[4] On *Calameo* which is a platform for digital magazines and on *Scribd* which is a platform for digital books. Cappelli is an ill fit for both.

A first step is a plain text version which found its way onto The-Colloseum.net. Then, the pages were indexed, and the index was made into a user-friendly interface. In fact, this was done by two different projects; one at Moscow State University, and one at Saint Louis University. Finally, we get to a fully digital database, where page numbers do not matter any longer, and ultimate flexibility is given for searching and interacting. This too has been done twice: one called *Cappelli online* (Zurich), the other *Abbreviationes* (Bochum).

The very last digital product referred to—*Abbreviationes*—has been in development since 1993 and is still undergoing maintenance. It thus outdates all the other digital products mentioned. It does not use any image from Cappelli's book as it even produces the graphical shape of the abbreviation by itself. Since it has also extended the corpus of abbreviations and their attestations, it has, in a way, grown into a thing of itself, a born-digital product. *Cappelli online* by Ad Fontes is a worthy competitor. With a much slicker interface and using small cut outs from the book to present the graphical shape of the abbreviations, it thus stays closer to Cappelli's original. Besides, *Cappelli online* is free, whereas *Abbreviationes* costs money.

Thus, with Cappelli's book as the standard reference work for abbreviations, many people in and around Latin paleography have found it irresistible to separate the entries from the page-structure and allow diverse ways of arriving at a specific entry—instead of the sole alphabetical order offered by the 1899 book—and thus make the contents of the work more accessible in the digital world and also more digital in nature. A similar path towards digital development can be traced for *Iconclass: an iconographic classification system*, which is a multi-volume classification system for art and iconography, originally done by Henry van der Waal, and also for *Die Wasserzeichenkartei Piccard im Hauptstaatsarchiv Stuttgart*, a multi-volume catalog of watermarks in ancient paper, originally done by Gerhard Piccard.[5]

Finally, paleography has also greatly benefited from digitized manuscripts in the production of facsimiles and albums of excerpts. Techniques have been developed to enhance the readability of images—for example, in cases of minor damage or bleeding through from the verso. A notebook describing both the technical process as well as an ethically responsible way of using it has been produced in the field of medieval music studies.[6] Retrieving the older text in

[5] Note that this is certainly not the only online watermark database. Watermarks, illustrations, and seals have been one of the most digitized (and electronically organized) subjects and treating them comprehensively would require a chapter on its own.

[6] Craig-McFeely, J., and A. Lock. *Digital Restoration Workbook*. Oxford: OSSC Publications, 2006.

a palimpsest can also be achieved by enhancing the digital photo by simply using open source software.[7]

1.2 *Born Digital, with A Hint of Print*

A plethora of websites created in the late 90's offer information about paleography (and manuscript studies more generally) in a style reminiscent of a printed publication. Notably, these pages often see a long time of steady updates and are still online as of today (June 23, 2018). A selection of such pages include Eric Voirin's *Cours de paléographie*, created in 1996 and last updated in 2014, Horst Enzensberger's *Buch- und Schriftwesen*, created in 1997 and last updated in 2013, Stephen Reimer's *Manuscript Studies: Medieval and Early Modern*, created in 1998, last updated 2015, Peter Doerling's *Sütterlinschrift lesen*, created in 1999, last updated in 2009, and Dianne Tillotson's *Medieval Writing*, created in 2000 and last updated in 2016.

The second wave of digital resources in paleography came online in the early 2000s. Examples include *Medieval Paleography on the Web*, developed between 2001 and 2005 by University of Leicester Centre for English Local History and the West Sussex Record Office. This website provides course materials for medieval and early modern English paleography. *Scottish Handwriting*, maintained between 2003 and 2015 by the National Records of Scotland, does something similar for early modern Scottish paleography. There is *Palaeographie Online*, run between 2003–2015 by a consortium of German academic institutions in Munich and Erlangen. It offers course materials for Latin paleography. From France, there is *Theleme*, launched in 2003 and still maintained by the École des chartes. This digital resource is actually a platform with various components. It has interactive facsimiles that can be used for (self-)training, proper courses, a French dictionary of abbreviations, and an extensive, systematic bibliography. A more modest website that offers a categorized list of links is the *Links für Handschriftenbearbeiter*, developed between 2003–2005 by the Kommission für Schrift- und Buchwesen des Mittelalter of the Austrian Academy of Sciences. In the same vein, we may note the creation of educational CDs and DVDs, such as by Evellum, which combined various media, including video, to deliver a self-contained, fully digital course on paleography and other aspects of manuscript studies.[8]

[7] Rafiyenko, D. "Tracing: A Graphical-Digital Method for Restoring Damaged Manuscripts." pp. 121–135 in *Kodikologie und Paläographie im digitalen Zeitalter 4*, edited by H. Busch, F. Fischer, and P. Sahle. Norderstedt: BoD, 2017.

[8] Muir, B.J. "Innovations in Analyzing Manuscript Images and Using Them in Digital Scholarly Publications." pp. 135–144 in *Kodikologie und Paläographie im digitalen Zeitalter*, edited by M. Rehbein, P. Sahle, and T. Schaßan. Norderstedt: BoD, 2009.

Compared like this, websites from the ’90s and 2000s make for a stark contrast. Whereas the websites of the ’90s were made by individual scholars and enthusiasts, and seem to be styled as though print publications, the websites of the 2000s are institutional and take better advantage of the possibilities digital technology offers. That is, that they do not wish to convey permanent knowledge in a page-by-page format offered for reading, but they ‘do’ something that requires regular updates and is given in the form of a list or is interactive. Bibliographies are an obvious genre to do this for, as the above websites testify. More specific for paleography is the rise of teaching resources. This is understandable when we look at the teaching materials available before the digital world. In the print world, instructors made a compendium of exemplary excerpts of a certain script that would allow students to see the particularities and slowly understand how to interpret a text written in that script. Creating these compendia involves a lot of work, and it is still only useful for a student if a teacher is present to point out what, exactly, is so special about each excerpt. In a digital environment, such compendia can be made quite quickly, and all kinds of extras can easily be inserted, such as circles, arrows, explanatory notes that hide/show upon a user’s request, and so on. Next to that, it is easy to allow a user to study an image of a manuscript, perhaps with some tools such as a magnifier glass or the option to rotate. The user can be given the chance to type out the text in a text box, and after clicking a button, the typed text can automatically be rated against the official typescript, and the differences can be dynamically shown.[9] In short, for the kind of didactics that paleography requires, a digital environment is very suitable. Peter Stokes observes that “if digital approaches should be closely integrated with and informed by humanities scholarship; then surely it follows that teaching should reflect this.”[10] It is interesting to notice how this has been taken up early on in the paleographies of different humanities subfields.

Of course, with aims set too high, an actual didactic tool might never be realized. For example, the University of London’s *InScribe* never made it, the *School* section on the website of the Hill Museum and Manuscript Library shows for years now a ‘coming soon’ message for most of its sections, and an educational platform called *Digistylus* has been announced and described, but

[9] Silke Kamp has an especially elucidating introduction on how to quickly make educational resources using simple, free software, see Kamp, S. “Handschriften Lesen Lernen Im Digitalen Zeitalter.” pp. 111–122 in *Kodikologie und Paläographie im digitalen Zeitalter*, edited by M. Rehbein, P. Sahle, and T. Schaßan. Norderstedt: BoD, 2009.

[10] Stokes, P. “Teaching Manuscripts in the Digital Age.” pp. 229–245 in *Kodikologie und Paläographie im digitalen Zeitalter 2*, edited by F. Fischer, Chr. Fritze, and G. Vogeler. Norderstedt: BoD, 2010, p. 232.

not made available.[11] As we shall see more in this chapter, teams of scholars frequently dream too big, thinking that by using a digital solution anything is possible, leading to numerous empty promises.

1.3 *Traditional Work Done Better*

A different line of computer support in paleography is the digital tools that are of help during the research phase. The most basic level is easily imagined but never discussed: paleographers open JPG images of manuscripts in an image editor, cut out letters or words of interest, and paste them in a Word document or a note-taking application such as Evernote or OneNote, to group together these cutouts in a systematic fashion so that a deeper analysis can be performed. When we look at higher level computer support, we come across what is called a 'virtual research environment.' I have listed below nineteen software applications that purport to be such VREs. This list is not meant to be exhaustive. I compiled it with an eye towards their applicability to manuscript studies and further wished to give a fair impression of the development trajectory and the scope of such specialized software.

The table lists on the left the preferred name or acronym. It then indicates where the production cycle currently is. Those marked red are unavailable or unusable, and there is no prospect in sight of resuscitation. Orange means the software is currently usable, but as there is no continued development, there is a potential of it becoming unusable or unavailable. Those marked yellow indicates that the applications work wonderfully right now, but there is no concrete, visible plan on the side of the developer(s) to continue to work on it, and so there is a danger of it becoming a 'finished' product. Finally, green indicates that the app developer is clearly invested in continuing to work on it.

In the next column, I have indicated whether the underlying source code can be viewed and used ('open source'), or whether it only allows you to use the application. A striped coloring means that only an older version and/or undocumented code was made available.

The last two columns indicate, roughly, the starting date of the development and the end date. All the ongoing projects also have an end date, because, on that end date, they delivered the fully functioning application, whereas

[11] Cartelli, A., and M. Palma. "Digistylus—An Online Information System for Palaeography Teaching and Research." pp. 123–134 in *Kodikologie und Paläographie im digitalen Zeitalter*, edited by M. Rehbein, P. Sahle, and T. Schaßan. Norderstedt: BoD, 2009. I see no substantial difference with another article: Cartelli, A. "DigiStylus: A Socio-Technical Approach to Teaching and Research in Paleography." pp. 741–753 in *Issues in Informing Science and Information Technology* 6 (2009).

TABLE 4.1 Comparison of Virtual Research Environments

Name	Status	Code	Begin date	End date
BAMBI	dead	closed	1995	1996
CEEC	out of date	closed	2001	2006
EPPT	dead	closed	2002	2005
TextGrid	slowed down	open	2006	2014
Agora	finished	closed	2006	2013
VRE-SDM	dead	closed	2007	2009
VMR CRE	slowed down	open	2008	2017
Aletheia	slowed down	closed	2008	2017
TextLab	slowed down	open	2008	2017
Teuchos	dead	closed	2009	2017
Graphoskop	finished	closed	2009	2009
SEASR	dead	closed	2010	2012
Scripto	revived	open	2010	2012
ArcheType	ongoing	open	2010	2017
Diptychon	dead	open	2011	2017
T-Pen	ongoing	open	2012	2016
Transkribus	ongoing	open	2013	2019
eCodicology	dead	closed	2013	2016
GraphManuscribble	not delivered	closed	2014	2016

'ongoing' merely means they are maintaining the application and/or developing the next version.

If we focus only on those which have slowed down or which are still ongoing, we may notice that they all do nearly the same: provide an environment to document manuscript evidence. They usually load an image, most of them then do automatic line detection, and then allow one to encode text and features and tag them to parts of the image. In most cases, it is possible to enrich the encoding with markup following the TEI standard, as explained in the next chapter. Often, the VREs offer the same or a similar interface for publishing finished transcripts, which can show the encoded texts and the image in an interactive manner. Notably, most of them can handle multiple users working on them simultaneously.

Their difference lies in the technology supporting them, the topic or purpose for which they were initially developed, and the way they are supposed to be used. Thus, TextGrid, VMR CRE, T-PEN, and Transkribus were all written mostly in Java, TextLab in Ruby on Rails, Scripto in PHP, ArcheType in Python, and Aletheia in C++.

TextGrid and *ArcheType* (formerly DigiPal) are the most versatile in intention, purposely built to support a wide array of humanities research. This may sound like a good quality, but it comes at the price of lacking certain functions that could have been of good use in your own particular project. Since they are both very big pieces of software, customizing them is not so easy. The *Virtual Manuscript Room Collaborative Research Environment* is also versatile, though specifically developed to support manuscript research in New Testament studies. Its suite of tools is seemingly as much about encoding and transcribing as it is about presenting the results. What is unique about VMR CRE is that it is entirely made up of pre-existing, open source components. Perhaps, this will make it easier to customize it and swap out one component for a better or newer one. *T-Pen* was also developed with premodern manuscripts in mind, such as the medieval 'Norman Anonymous.'[12] T-Pen has line detection and actually centers its transcription tools around the line as a unit of measure rather than a full page. *Aletheia* and *Transkribus* (formerly tranScriptorium) are geared towards automated optical character recognition, sometimes in a manuscript context called handwriting recognition (HWR). Aletheia does a particularly good job at extracting all features and characters on a page and seems to have been made with an eye towards ancient manuscripts. It can, therefore, be used very well for detailed paleographic research. Transkribus, on the other hand, shines new light on large bodies of texts from the early modern period, especially when worked on in a team or even crowdsourced. It thus draws its strength not from the attention to detail but from drawing the wider picture. *TextLab* and *Scripto*, finally, were both developed to support the diplomatic transcriptions of early modern archival materials. TextLab did so for a project on the personal archive of the novelist Herman Melville, and Scripto did so for the archive of the 18th century War Department. TextLab's project is remarkable in that Melville's papers show an incredible landscape of writing, rewriting, inserting, crossing out, and all of this on whatever piece of paper Melville could get his hands on. TextLab tries to sort out this chaos,

[12] This tract is preserved in Cambridge, Corpus Christi College, MS 415, and was edited using T-Pen as the Electronic Norman Anonymous Project. Tellingly, despite the $200k grant in 2008, the digital edition can no longer be accessed.

which serves as the basis for all kinds of research, such as new insights into his writing process. Scripto's original project was perhaps even more innovative. The archive of the US War Department of the last quarter of the 18th century was lost in a fire. Copies of many of the documents can be found in all kinds of other archives, and so this project attempts to digitally bring together all these documents to reconstruct the War Department archive.

The different purpose of each VRE translates into different preferred workflows. TextGrid, T-Pen, and Transkribus require you to login and store your documents on their servers. This means you can only use them effectively when you are connected to the web. ArcheType can be downloaded and run locally, but this requires significant computer skills. All four come with a noticeable learning curve, for which some online documentation is provided. I could not test VMR CRE and TextLab, as I could not get them running.[13] They require more time to learn of the technical aspects than I was willing to invest. Similarly, I could not try all the functions of Aletheia since it is a paid software. It is a polished program with the look and feel of a Microsoft Office product. It, therefore, has only a small learning curve. Scripto, finally, is a bare-bones approach to getting a tool online quickly, capable of accommodating both basic transcriptions as well as providing a way to display transcriptions that are done. To do this, you need to have your own website running, preferably with an Omeka content management system. This is easy enough to do, even if this will be your first time doing it. With this ease, there are dangers. For example, Scripto was supported for Wordpress and Drupal too, but this relied on an external service that, at the time of writing, was not functioning. It is good, then, that Scripto secured new funding for an overhaul and update.

It is quite shocking how the majority of VREs has kept their source code closed or only shares parts of it, such as older or undocumented versions. Similarly, it is miserable to notice that half of the VREs are no longer alive and that many others are on their last legs. The basic websites from the 90s we noticed earlier in this chapter perform better than these VREs. From the funding details of a number of them, it is clear that this table represents millions and millions of Euros and Dollars. TextLab, Scripto, and T-Pen are North American projects funded in large part by the NEH. The others are European-funded, both by national agencies such as DFG and international ones such as EU's Horizon 2020 fund. From the dates, it is clear that we see the rise of big projects funded by grant agencies only in the late 2000s. I have included one outlier

[13] VMR CRE came out of doctoral research by Troy Griffitts. His dissertation makes for excellent reading: Griffitts, T.A. "Software for the Collaborative Editing of the Greek New Testament." PhD dissertation, University of Birmingham, 2017.

in this regard, known as *Better Access to Manuscripts and Browsing of Images*, which existed in the 90s. It has left no trace of its existence other than a discussion in some publications,[14] but it does show that such digital approaches to manuscript studies were a long time coming.

You might wonder if you should use any of them. Two important conclusions we can draw from this brief comparison is that one size does not fit all, and that these pieces of software age quickly and badly. We can only make use of them insofar as we test that they truly help us achieve our goals, and as long as we keep in mind that they will die someday.

1.4 *Towards Handwriting Recognition*

Computer technology can sometimes provide new opportunities. One simple example is that with the switch from the print world to the digital world, scholars have developed typography that is much more extensive than was available for print, that can do much more justice to the intricate writing systems of previous centuries.

Another simple, yet brilliant, idea is the use of *Zooniverse*, a website that allows scholars to set up a project for what they call 'citizen science' but what is more commonly known as 'crowdsourcing.' The projects on *Zooniverse* are very big and attract a large group of people who volunteer their time. Paleography, and manuscripts studies in general, can also benefit from this, as has been most successfully proven by the "Scribes of the Cairo Geniza" project.[15] The general public has already classified tens of thousands of snippets coming from this medieval 'sacred trash' according to generally easy to identify characteristics such as the language of the script and whether the script is formal or informal. Not only is this form of data collection immensely innovative, but it will also get us a data set of proportions that would otherwise be impossible to achieve.[16]

[14] Bozzi, A., and S. Calabretto. "The Digital Library and Computational Philology: The BAMBI Project." pp. 269–85 in *Research and Advanced Technology for Digital Libraries*, edited by C. Peters and C. Thanos. Berlin: Springer, 1997; Calabretto, S., and A. Bozzi. "The Philological Workstation BAMBI (Better Access to Manuscripts and Browsing of Images)." *Journal of Digital Information* 1, no. 3 (1998); Babeu, A. "Rome Wasn't Digitized in a Day": Building a Cyberinfrastructure for Digital Classics. Washington: Council on Library and Information Resources, 2011, pp. 158–160.

[15] Eckstein, L.N. "Of Scribes and Scripts: Citizen Science and the Cairo Geniza." pp. 208–214 in *Manuscript Studies* 3, no. 1 (2018).

[16] A different from of crowdsourcing is by pulling in all available scholarly literary and automatically analyzing their contents, thereby establishing keywords for each fragment. This has been reasonably successful in the case of, again, the Cairo Geniza, see Stokoe, Chr., G. Ferrario, and M. Outhwaite. "In the Shadow of Goitein: Text Mining the Cairo Genizah." pp. 29–34 in *Manuscript Cultures* 7 (2013).

Similarly, simple yet innovative is the online tool *RetroReveal*,[17] where paleographers with no knowledge of complicated image manipulation software can upload a photo of interest and the website will return it with all kinds of different manipulations performed on it. When working with, for example, faded script, palimpsests, or seals, this can be a real boon.

A much more advanced use of computers is the attempt to find a script's characteristic shape, for example, to search through a large pile of snippets and find snippets and folia that ought to have belonged to the same manuscript.[18] There are also various ways to analyze the shape of the script to ascertain the particularity of it specific to one person, for example, by analyzing the TEI-compliant transcription of the text.[19] Another possibility is by taking measurements of certain aspects of the codex.[20] A third way is by an automated investigation of abbreviations.[21] Finally, such typologies can also be done by the raw analysis of the shapes the ink makes on the paper, what Mark Aussems calls a 'scribal fingerprint.'[22] Slightly more

[17] Erickson, H.M., and J. Ogburn. "RetroReveal.Org: Semi-Automated Open-Source Algorithms and Crowdsourcing Tools for the Discovery, Characterization and Recovery of Lost or Obscured Content." p. 80 in *Archiving 2012*. Copenhagen: Society for Imaging Science & Technology, 2012.

[18] Wolf, L., N. Dershowitz, L. Potikha, T. German, R. Shweka, and Y. Choueka. "Automatic Palaeographic Exploration of Genizah Manuscripts." pp. 157–179 in *Kodikologie und Paläographie im digitalen Zeitalter 2*, edited by F. Fischer, Chr. Fritze, and G. Vogeler. Norderstedt: BoD, 2010.

[19] Stapel, R. "The Development of a Medieval Scribe." pp. 67–86 in *Kodikologie und Paläographie im digitalen Zeitalter 3*, edited by B. Assmann, J. Puhl, and P. Sahle. Norderstedt: BoD, 2015; McGillivray, M. "Statistical Analysis of Digital Paleographic Data: What Can It Tell Us?" pp. 47–60 in *TEXT Technology* 1 (2005); Driscoll, H. "The Legendary Legacy: Crunching 600 Years of Saga Manuscript Data." pp. 71–79 in *Kodikologie Und Paläographie Im Digitalen Zeitalter 4*, edited by H. Busch, F. Fischer, and P. Sahle. Norderstedt: BoD, 2017; Stutzmann, D. "Paléographie Statistique Pour Décrire, Identifier, Dater … Normaliser Pour Coopérer et Aller plus Loin ?" pp. 247–277 in *Kodikologie Und Paläographie Im Digitalen Zeitalter 2*, edited by F. Fischer, Chr. Fritze, and G. Vogeler. Norderstedt: BoD, 2010.

[20] E.g. Brey, A., and E. Muhanna. "Quantifying the Quran." pp. 151–173 in *The Digital Humanities and Islamic & Middle East Studies*. Berlin: De Gruyter, 2016.

[21] Gottfried, B., M. Wegner, M. Spano, and M. Lawo. "Abbreviations in Medieval Latin Handwriting." pp. 3–9 in *Manuscript Cultures* 7 (2013).

[22] Aussems, M., and A. Brink. "Digital Palaeography." pp. 293–308 in *Kodikologie und Paläographie im digitalen Zeitalter*, edited by M. Rehbein, P. Sahle, and T. Schaßan. Norderstedt: BoD, 2009; Stokes, P. "Palaeography and Image-Processing: Some Solutions and Problems." *Digital Medievalist* 3 (2007); Herzog, R., A. Solth, and B. Neumann. "Computer-Based Stroke Extraction in Historical Manuscripts." pp. 14–24 in *Manuscript Cultures* 3 (2010); Herzog, R., A. Solth, and B. Neumann. "Computer Methods for Comparing the Hands of Manuscripts." pp. 169–175 in *Manuscript Cultures* 4 (2011).

general is an automatic assessment of the script to date and place the manuscript.[23]

The crown jewel of digital manuscript studies, automatically converting the text that manuscripts contain from a digital photo into plain text, requires a longer discussion. Such conversion would make it possible to abstract the text itself from the artifact, allowing it to be searchable. This, in turn, will allow us to automatically store the text in a file or a database, from where we can construct automatic links between different texts. One may imagine that if text recognition would be flawless, a critical edition could be made automatically by letting the computer read through the different manuscripts. After setting up rules for the computer to figure out how to combine the different versions of the text as presented in the different manuscript, the computer could turn those texts into an interactive, digital edition, or into a print edition according to a certain critical approach, including an automated generation of a critical apparatus.

However, reaching that ideal state is far from trivial. For instance, what is 'the text' referred to in my second sentence? Manuscripts rarely keep the text as a uniform body such as a printed book nearly always does. Words can be jammed in, above or below a line, as a final addition. There can be all kinds of paratexts, some of them important for a correct reading of the main text (e.g., corrections), some of them important for a correct understanding of the text (e.g., marginal comments). Even provided that the computer can accurately read all texts, it is still not obvious how all of them can be stored in a way that faithfully represents the manuscript.

Prior to this problem of faithful representation is simply the problem of optical character recognition itself. Contemporary handwriting recognition has been developed fairly well. For example, companies and governments that collect information from large groups of people by paper forms—the government's tax forms, for example—rely on the computerized reading of them. Such problems are simplified by offering only limited options, and by persuading users to write each letter in capital in a separate box. There are more advanced examples, as well. Police and intelligence agencies, for example, need to analyze vast quantities of handwritten texts for suspicious contents but also to analyze the style of handwriting—that is, to determine whether two pieces of writing are from the same hand. This kind of handwriting recognition has spawned a small industry dedicated to this problem, using

[23] Christlein, V., M. Gropp, and A. Maier. "Automatic Dating of Historical Documents." pp. 151–164 in *Kodikologie und Paläographie im digitalen Zeitalter 4*, edited by H. Busch, F. Fischer, and P. Sahle. Norderstedt: BoD, 2017.

fancy terms such as neural networks and hidden markov models.[24] The visibility of this industry and its first successes have perhaps given the impression that the problem of optical character recognition has been solved. But this is not so—at least not for paleography in the true sense of the word: the study of ancient writings. Porting the solutions for today's handwritings to ancient handwritings is not straightforward, with several factors playing obvious roles. For one, today's handwriting, for most languages, is more and more letter-based instead of word-based, indicating that letters are written separately, even if they are nominally connected. Since modern OCR technology uses this to compartmentalize words into letters, the same technology cannot be applied to older and more connected scripts. A related factor is that, in the past, much more ligatures were in use. Similarly, much more abbreviations were in use, sometimes using complex notations that are easy for a hand to write and for eyes to read, but difficult to detect based on the sheer pattern of ink on paper. Another factor is that the corpus of ancient writings is much, much smaller than contemporary writings, thus making it harder to train a computer. Quite often, for the most interesting texts, there is only one manuscript witness, and its handwriting will have unique features not seen in other manuscripts. Whereas human eyes and a strong erudition can relate those features back to other scripts to figure out the meaning, computers will be at a loss to understand the blobs of ink. Furthermore, a fair few manuscripts—it seems especially those of interest—are damaged or otherwise hard to read. Whereas contemporary handwriting is often black or blue ink on pristine white paper, often with pre-printed lines to guide the sentences, manuscripts will be brownish or yellowish, with specks here and there; the ink may be faded or bleeding through the back-page, and tears, rips, and wormholes could show up in a black and white image. The script may be irregular or written at different angles. Again, whereas human eyes and erudition can separate out those imperfections from the text—often based on semantics—a computer will find it nearly impossible to decide which black spot is garbage and which black spot is valuable. On top of this, the literature on handwriting recognition hints that they use very high-resolution images, but this is simply not the reality of many of our fields, as I point out in Chapter Three.

Despite all these obstacles, steady progress is made in filtering out the script from digital photos of manuscripts. A survey of the state-of-the-art OCR technology for ancient manuscripts seems futile for three reasons. First, such a survey would have to be very long to take into account the work done on all kinds of scripts and periods. Indeed, virtually any disciplinary subfield of manuscript

[24] Several academic journals specialize in it, such as Elsevier's *Pattern Recognition Letters*.

studies has seen some progress towards this goal, even if only as a proof of concept. Second, such a survey would be highly technical and, therefore, impenetrable for those who did not major in Math or Computer Science. This is because most publications on OCR technology are brief statements on the success rate of applying one or other advanced methods and do not discuss the practical implementation for humanities research purposes, but focus on the technical achievements. For lone students and scholars of humanities, this is not that interesting. Third, this industry is, at the time of writing, moving and changing fast. Thus, a survey would become quickly outdated.

That leaves us to discuss what benefits we can reap from automated handwriting recognition. Such technology often relies on 'training' a computer, so that it can build a corpus of shapes for which we manually instruct them which letters (or words) they correspond to. This is done by first typing out several thousands of words. If you have five manuscript copies of the same text, you may need to do this up to five times, if the handwriting is too different for the computer to recognize. The obvious case for which this would be a fine investment is if you are researching an archive of unique documents in the same handwriting, such as a log, chancellery books, a diary, or correspondence. In that case, the corpus is vastly larger than the amount of manual typing you need to do, and you most likely want to use text recognition mostly to find passages of interest. Additionally, those writings often have a normalized *mise-en-page*, with parts of the page always reserved for a title, a date, an amount, or a place name. Such regularities can be exploited to point the computer towards regions of interest and store the extracted information in the correct way. The case is very different for the study of premodern texts. With most texts of intellectual or cultural value, you wish to do a close reading and critical editing of the text, and for that, you will often use a variety of different manuscripts. These manuscripts are likely to have nothing in common towards the specific hand they are written in, nor is their *mise-en-page* similar (or even regular). Setting up automated text recognition might be too laborious in that case. I imagine that the first texts outside archival materials to take advantage of such technology will be scientific texts with their small vocabulary (like 'square' or 'plus') repeated over and over again, making it easy to train a computer to read such texts, and also poetry, which often has a clean and regular *mise-en-page*. For the former, the challenge to overcome is that scientific texts can be written in a sloppier hand and more crammed on a page (since the market for such texts was only tiny), making it harder for a computer to lock on and identify words. For the latter, the challenge to overcome is that poetry can have extensive and obscure vocabulary, making it harder for a computer to match the pattern with a word.

Then again, even if the premodern text is huge and you gladly train the computer in four different handwritings, it seems unlikely to me that automated text recognition is going to do much good. The manuscript world has a different concept of what a text is, as I explained in Chapter One, and this means that even the internal comparison of body-text and paratexts can easily become too atypical to be properly caught by an algorithm. Mostly, 95% of the time, a text is simply what is in the body of the folio, word for word, but it is exactly in the 5%, where something unusual is going on—an emendation in between the lines, an addition in the margin, a large omission in one manuscript—that the editor can add value by doing the analytical work for the reader and laying out the different textual elements as they ought to be.

A similar problem manifests itself on the word level. The 99% of a manuscript which is legible is not the problem; whether a person or a machine does them is merely a difference of execution time, but those 99% is hardly the value that a critical edition brings. It is the 1% which is difficult to read, where the pen of the scribe might have slipped or where a reader might have made a hyper-correction with ink just a tiny bit different from the original ink—in those cases, the erudition of an editor can weigh in to make the right decision. Meanwhile, the computer, in the words of Craig Baker, "remains incapable of distinguishing between a minor variant and a significant error, and has thus not brought with it any tangible methodological advantage in this respect."[25] Baker writes this in 2010, but as of 2019, I would still support this statement. This comes back to the problem of paleography noted at the beginning of the chapter, whether paleography is a science with hard, measurable arguments to make, or an art giving more weight to soft, erudite arguments. Even the simple paleographic task of transcription is not that easy. Do we expand abbreviations? Silently correct the gender of a verb? Use a modern way of writing a letter? Do we leave out crossed out words? Indicate rubrics? We may find that for the same question, we answer differently, judging by the context, and the purpose of our transcription. This contextualization counts even more for the letters themselves. For example, in the word 'learning,' we recognize the middle two letters as 'r' and 'n,' but in the word 'glearning' you might have mistakenly assumed there was an 'm' in the middle. In the digital world, this is more or less easy to spot,[26] but in the manuscript world, it is not easy at all and happens very frequently. The context will help you decide what the letters should

[25] Baker, C. "Editing Medieval Texts." vol. 1, pp. 427–450 in *Handbook of Medieval Studies: Terms—Methods—Trends*, edited by A. Classen, 3 vols. Berlin: De Gruyter, 2010, p. 440.

[26] Although, it may be noted, this exact technique is used for phishing to give the impression a message is coming from a credible source.

be. Since such decision making is at the heart of paleography, I conclude that paleography is not hard enough to be fully automatized, but not soft enough to disregard automated analysis. I think, then, that automated text recognition, when it is advanced enough to be implemented into our normal workflow, will be a great help for editors, but it remains only that—a help.

2 Rise and Fall of Team Projects Funded by Grants

I have sketched a narrative where paleography, from early on, has eagerly engaged with computer-supported solutions and, fueled by big grants, become increasingly project-centered. This has resulted in a widening gap between the lone student or scholar and those who develop electronic tools. We need to find out about the existence of such tools, then we need to get it functioning, then we need to learn how to use it, and only then can we start to do our work in it. With most scholars and even most students (despite having grown up in the digital world; born-digital, if you will) not being tech-savvy, adoption of such tools has been low. Big-grant projects that produce such tools have noticed this too, and the most common suggestion is to foster a community that rallies around a tool.[27]

On top of that high barrier to entry, we run the risk of the software malfunctioning or being entirely shut down, without the guarantee of any support. Big influx of cash for building a tool also means that there will be a point in time when there will be a huge drop in funding, quite often, in fact, it goes from all to nothing. Yet, computer technology does not age well without maintenance. In this aspect of continued usability, the making of digital tools is very different from print publications. Whereas a book is self-sufficient (you only need to know the alphabet the book is typeset in and the language it is written in), all kinds of extraneous factors—the hardware and the operating system—need to be just right for software to work. When a tool can only be used remotely over the internet, then, when the project stops paying for the server on which the application resides, the app will simply stop existing. This puts the entire product at jeopardy and, therefore, makes using it, even if the tool is currently flourishing, a lot less attractive. Will the software store and output our results in a way that will be meaningful now and in the long term? If a critical

[27] As much is stated in a report entitled "SEASR—Software Environment for the Advancement of Scholarly Research", and also by TextGrid see Neuroth, H., A. Rapp, and S. Söring, eds. *TextGrid: Von Der Community—Für Die Community*. Glückstadt: Verlag Werner Hülsbusch, 2015, p. 33.

aspect breaks, who will fix it? All too often, big-grant projects do not provide documentation that answers these questions. This leads me to conclude that, from the point of view of the uncertainty of continued, paid development and maintenance, even the short-term future of software is precarious.

A hidden assumption in the previous paragraph is that the source code of the software is openly available and licensed in a way that allows free transformations (fixes and updates), leaving the option open of free (voluntary) maintenance by people outside the original team. This, however, is not as straightforward as it may sound. First, we should note that if a tool has been put together by a team, it will likely be of a complexity similar to the team structure, indicating that it is difficult for an individual outside the team to fix or amend it.[28] Second, and more importantly, a common aspect of big-grant, team-effort tools is the unwillingness (or carelessness?) of sharing the source code. This makes the software a fossil once it is abandoned by the big-grant team. As we can surmise from the discussion above, closed source software is widespread among big-grant projects, and I think it is a big reason for the unpopularity and eventual abandonment of a tool.

Not disclosing the source code is problematic for more and bigger reasons than the third-party impediment of future development. There is, I believe, an ethical component at play here. For projects that have run on public funds, it seems simply indecent to not give back to society. If we ask scholars to execute a project on behalf of the society, it should not be that only those scholars reap the benefits of it. That would not be fair to society, and it would also give an unfair advantage to the scholars who do have access to it. A similar discussion on open access publishing has been ongoing for many years, and the tide is slowly turning towards open access. In a similar vein, humanities software should become open source by default.

A result of not disclosing the code—in most cases not even discussing it—is that the software becomes a 'black box,' a 'magic device' in which anything can happen without any assurance of its veracity.[29] In more scientific terms, the experiment the code can execute cannot be replicated and, therefore, not verified. In that case, using such software fails to meet well-established academic

[28] I am referring here to Conway's Law which states that "organizations which design systems [...] are constrained to produce designs which are copies of the communication structures of these organizations." Conway, M.E. "How Do Committees Invent?" pp. 28–31 in *Datamation* 14, no. 5 (1968).

[29] This has been noted before, see Hassner et al., "Computation and Palaeography: Potentials and Limits," p. 6; Stokes, "Computer-Aided Palaeography, Present and Future," p. 31; Solth, A., R. Herzog, and M. Neumann. "A Modular Workbench for Manuscript Analysis." pp. 132–137 in *Manuscript Cultures* 7 (2013), p. 133.

criteria. This is a particularly painful point as it is easy to suppose that the big-grant projects are well-meaning and merely want to deliver a polished user experience. The commercial world of consumer hardware and software has rapidly moved that way. Examples of hardware include Amazon's e-reader, the Kindle, and Apple's tablet, the iPad, whose inner workings are sealed off and whose tactility gives it a direct, almost thoughtless, user interaction.[30] With software it is the same—for example, the technology that makes social media applications run smoothly, such as Facebook's React which makes automatic updating in the browser possible so you can keep scrolling down infinitely and have autocorrect and autocomplete, or Twitter's Bootstrap, which makes websites render well regardless of the medium through which you are looking it up, be it a phone, tablet, or computer. The big tech companies promote this paradigm aggressively in their marketing. Big-grant developed tools follow this tone, by using declarative statements about the capacities of the software, thereby signaling that these capacities are not to be questioned, and talking in the passive voice when speaking of the development of the software as though its development has been devoid of any human decision.[31]

A corollary problem is that the more polished an application is, and the more general its purpose is stated, the less it will be suited or adaptable for a researcher's specific purpose. In the humanities, rarely can two projects be done the same way or two sources be studied and analyzed uniformly. Using a digital tool of a big-grant project requires doing your analysis less precise as—more likely than not—the tool has not been developed for your use case. This is a cost you incur that is hard to counteract.[32]

Since the end-user is now completely at the mercy of the team, this problem is aggravated when the tool demands a user to upload his or her material to the team's server. The server might be terribly slow or sometimes go offline, and if next to the raw data also notes and analysis are saved on that server, the scholar is no longer in control of their own work. Next to questions of continuity and technical compatibility, new questions concerning legality and ethics now pop up. For example: what if we paid a library for photos but the terms of the library state they can only be used for individual research purposes? Are we

[30] Merkosi, J. *Burning the Page: The Ebook Revolution and the Future of Reading*. Naperville: Sourcebooks, 2013, p. xvi; Emerson, L. *Reading Writing Interfaces: From the Digital to the Bookbound*. Minneapolis: University of Minnesota Press, 2014, p. 24; McLaughlin, T. *Reading and the Body*. New York: Palgrave Macmillan, 2015, p. 179.

[31] I feel it would be undeserved to cite specific people or projects.

[32] Smith, N. "Digital Infrastructure and the Homer Multitext Project." pp. 121–38 in *Digital Research in the Study of Classical Antiquity*, edited by G. Bodard and S. Mahony. Farnham: Ashgate, 2010, p. 136.

allowed, in a legal sense, to upload these photos to a server of a big-grant tool? By uploading, does the big-grant team now own those photos and are they allowed to do other things to it? What if the library, the user, and the server are in different countries? Do different laws apply? What if the work of a scholar on such a server-hosted tool is deleted without notice? Has the scholar some rights to get their work back or appeal the decision? So far, scholars (and students) in the humanities have operated on good faith when it comes to these legal issues. That has worked well in as much as scholars mostly worked on their own computers, but once a server-hosted tool is involved, the user will be exposed to much more imminent legal issues. Next to this are ethical concerns. If the archival material contains private data, such as personal details that would not be incorporated in the final analysis, it may not be alright to leave that on a server you do not control yourself. In short, server-hosted tools needlessly complicate our workflow, exposing us to significant liabilities.

So far, we have discussed the drawbacks of such tools. Let us now focus more on the team-aspect inherent to developing a complicated tool at once. The very notion of 'team' in the sense of working together towards one product is quite foreign to many humanities disciplines. The norm is lone scholars working on a project by themselves from start to finish, resulting in a publication with only one author (themselves). Of course, throughout the project, ideas and drafts are bounced off of colleagues. Conferences and teaching duties are other parts of a scholar's normal life in which they can contribute to projects of others and receive contributions, mostly through criticism and pointing out relevant primary sources and secondary literature. But throughout this process it remains the work of the one scholar, and as a result we rarely encounter multi-authored publications. Even in digital humanities, when it is time to publish, single-authored papers are the norm.[33]

For big-grant tools, the building of a tool has virtually always been proposed as a joint effort by humanities scholars, who think of what the tool should do, and engineers who think of how technology can do that.[34] Some scholars boldly state that "high level interdisciplinary collaboration between humanist research and computer science was *demanded*" (emphasis added).[35] Sometimes,

[33] Nyhan, J. "Joint and Multi-Authored Publication Patterns in the Digital Humanities." pp. 387–399 in *Literary and Linguistic Computing* 29, no. 3 (2014).

[34] Bradley, J. "No Job for Techies: Technical Contributions to Research in the Digital Humanities." pp. 11–25 in *Collaborative Research in the Digital Humanities*, edited by M. Deegan and W. McCarty. London: Routledge, 2012.

[35] Busch, H., and S. Chandna. "ECodicology: The Computer and the Mediaeval Library." pp. 3–23 in *Kodikologie Und Paläographie Im Digitalen Zeitalter 4*, edited by H. Busch, F. Fischer, and P. Sahle. Norderstedt: BoD, 2017, p. 6.

teams claim this has worked remarkably well.[36] But the majority of cases run into difficulties. The joint statement of a seminar on digital paleography in Leibniz is brutally honest. Their number one finding is that "difficulties in communication between palaeographers and computer scientists is a prevailing problem."[37] It is worth quoting in full a later statement:[38]

> It might seem at first that problems in communication are easy to solve, and that it is "just" a matter of listening and understanding, a matter of ironing out differences. However, even in our group of twenty people at Dagstuhl from different backgrounds, where all were accustomed to collaborative scholarship, a striking recurring difficulty in understanding each other was apparent.

Two points are made that are especially interesting to us. First, even at a high-level meeting such as this one, where members of both parties were experienced in the topic at hand and motivated to discuss it with the other, communication broke down. Second, they emphasize that they were all accustomed to working with colleagues, and so the problem cannot be a lack of social skills, but it must be sought in their difference of expertise or education towards that expertise. Peter Stokes echoes this sentiment, pointing out that it is all fair and well to say that a humanities scholar is not expected to understand "the intricacies of postgraduate-level mathematics," but this does mean that "if we cannot understand them then we cannot evaluate them properly or debate their results."[39]

2.1 *Archetypes across the DH Spectrum*

As a solution, the people from Leipzig suggest that a new breed of academics is necessary: "a middle-person, a translator: a person who is versed enough in each of the collaborating fields to understand enough of each of the discipline-specific lexical fields to foster good communication and fruitful exchanges."[40] Stokes, on the other hand, sees more success in "a 'lone scholar' working on all aspects of the topic, theoretical and practical, 'digital' and 'humanities.'"[41] Of course, these solutions do not have to exclude each other. Such a lone, well-

[36] See e.g. Muir, who says "I feel that it is what has given us the edge over our colleagues during the past decade." p. 137.

[37] Hassner et al., "Computation and Palaeography: Potentials and Limits," p. 2.

[38] Ibid., p. 13.

[39] Stokes, "Computer-Aided Palaeography," p. 322.

[40] Hassner et al., "Computation and Palaeography: Potentials and Limits", p. 16.

[41] Stokes, "Computer-Aided Palaeography," p. 326.

versed scholar would be an excellent candidate as a middle-person. More importantly, we do not have to think of it as a binary choice with now a third option in the middle. Computer-supported research does not need to be seen as 'digital humanities', cut off from 'classical humanities.' Instead, I think, we are better off if we place such research on a DH spectrum. To get a better grip on such a spectrum, I propose to introduce six archetypes in which the vast majority of humanities scholars and students can be categorized.

The Believer. On the utmost right side of the spectrum, we find *the believer*, somebody who is fully absorbed in the digital humanities. They have made it their field in which they want to make a career. Hence, their research is mostly geared towards developing and applying new technologies, pushing the boundaries of what is possible. They have advanced and intimate knowledge of programming and the way technology works, since their study or personal development time is maximally spent on it. They see no great problem in the perceived chasm between 'classical humanities' and 'digital humanities' and would likely not agree that the team-based, big-grant projects are a failure.

The Obstinate Ostrich. On the other side of the spectrum we find *the obstinate ostrich*. This archetype accommodates a diverse group of people in the humanities who ignore computer technology not out of ignorance but as a choice and so they too are perfectly happy with seeing 'digital humanities' as something entirely different from 'classical humanities.' Essentially, they won't go beyond using the computer as a typewriter. As a consequence, they feel it is unfair that what they can accomplish by great labor, some people now accomplish with a simple search through a database. If we try to grab their attention we run the risk of being repaid with a very negative response. Since this archetype is discernible among people in influential positions, this is not to be overlooked. They can and will shut down "this DH nonsense" if we are not careful.

The Sour One. Going back to the right side of the spectrum, just left of *the believer*, we find *the sour one*. They did not want to showcase their technological progress only to people in DH, but they wanted to have it accepted by their peers. They have, by all accounts, already tried to be that middle man, but for a complication of reasons, this has so far fallen on deaf ears. There is a danger for people using tech in the humanities to become this archetype. Perhaps a way to avoid it is to choose carefully who you engage with; both *believers* and *ostriches* are not ideal conversation partners if the initiative for collaboration did not come from them.

The Spider. Not too far off from *the sour one*, we can find *the spider*. They are usually professors at the helm of a research team that is doing exciting stuff with technology in the humanities. As a team effort, their tech level falls at the mid to high level of the spectrum, but *the spider* him- or herself usually only has passive knowledge of the possibilities and limitations of technology. Their added value lies primarily in having a large network and the ability to attract grants. Through their grants they can employ people who *are* capable of wiring together tech and humanities research. Their main job, then, is to connect people. In strengthening the entire spectrum of DH, and making computer technology a normal part of humanities research, they are very important.

The Blind and the Lame. Here we have a symbiosis between a professor who is enthusiastic about digital solutions, and a student or hired professional who can get the technology to work. Often times, the professor, relying on their expertise, will come up with research questions that are closer to classical humanities than research questions a *believer* or *spider* would come up with. The professor sits on top of the shoulders of the student, instructing them to go this or that direction. I think actually that this archetype is not very helpful. It would be better if we can redirect the enthusiasm of the professor by for example giving the student more opportunities to open their eyes and learn to walk by themselves.

The Centaur. It is very likely that you, as a reader of this book, are a *centaur*. *Centaurs* are students and scholars who are simply working in the humanities but are devoting real time to learn how to use computers in a serious way. Some spend more time on this than others, but for all counts that their head is firmly rooted in conventual methods and practices of their field, while their feet are taking on distinctly digital shapes. As long as they do not let their heads also turn digital, which would turn them in *believers*, this group has the greatest potential to fill that spectrum and decisively shift the practices of the humanities towards the digital world in which, to be honest, we already live. Because *centaurs* obtain an active skillset in using computers in their workflow, and because they do so in the context of their own study or research, they are self-sufficient and do not require other people or grants to implement their computer skills.

3 Drawing Ancient Symbols on a Tablet

After getting to know these very expensive projects that produced complicated software, and if you are ready to be a centaur, let us start with something simple yet effective. Mastering a basic skill, such as recreating symbols of paleographic interest on a computer, will be a major boon to manipulate and analyze them easier. In this section, we will discuss how to do that. As with most chapters in this book, our discussion will be shaped by a case study, to get a better sense of how it can be of practical benefit. The case study was done with an expensive tablet and a stylus, as the haptic interaction provides a uniquely fine experience to do this kind of work. I am assuming you already have such a device for general consumer needs. I use an iPad Pro from 2017 (model A1701), but what I explain here is device-independent and can be done on an older or a newer tablet. Moreover, much of the procedure described here can be done on a normal computer too, with a mouse.

We will learn to use a function that will remain a feature of all vector-based drawing applications, namely the pen tool. I use the app Vectornator, and on my computer I use Adobe Illustrator. The latter is a commercial, paid software. You may obtain it through your university or look for a free alternative, such as InkScape, which can run on Windows, macOS, and Linux. For a drawing application to be vector-based means that everything you draw is not stored as color values for pixels, but as mathematical points, lines, and shapes. This means that you can zoom in all you want but nothing will look pixelated because the curve of the lines and shapes will be recalculated and redrawn. With the pen tool, you can add points (often called anchor or corner points) on a drawing area (usually called canvas) to make a line. And if you make the last point coincide with the first, it will recognize the entire line as a shape (often called a path). What is especially attractive about the pen tool is that, at each point, you will have two levers or handles on each side of the point relative to the line it is a part of. And by selecting and dragging these levers, you can adjust the curve the line makes on that side of the point. Lastly, the pen tool helps separate the different tasks of drawing a line and/or shape and giving that line and/or shape an appearance: first, you draw the contours, and only then do you select a particular line and fill the style—for example, whether you want an outline or not and whether that outline should be dashed or solid, whether the shape should have a gradient fill and where that gradient should start and end, and so on.

Thus, using the pen tool is as simple as first adding points that roughly outline the shape you want to draw, then adjusting the curvature of all the segments of the outline to perfectly match the shape you want, and finally to give the shape the colorful appearance you want. Shapes with holes in them can also be drawn by drawing the holes first and giving them a different color, then the outer shape, selecting all, and using the function to 'combine paths' (the icons should guide you to the right choice). It is called 'Path' in Vectornator and 'Pathfinder' in Illustrator.

To understand the intricacies, we will walk through a project from start to finish: namely, a better understanding of three symbols a medieval Islamic philosopher wrote which, he says, represent the essence of his teachings.[42]

[42] I published my findings before and parts of that article are reproduced here: Lit, L.W.C. van. "Mysterious Symbols in Islamic Philosophy." pp. 34–39 in *Islamic World of Art* 3 (2017).

To find symbols in a philosophical text is surprising. Manuscript copies of Islamic philosophical texts consist of walls of text, page after page. Readers of Islamic philosophy were not interested in embellishments or illustrations. Neither were writers; only very seldom did they make use of graphics or symbols to get their point across. Suhrawardī (d. 1191) is such an exception. Even though this philosopher only lived to be 36 years old, he produced an extraordinary philosophical output in which he advanced a great number of innovations. He himself was wont to describe these innovations in terms of an entirely new system of thought, which he dressed up in a vocabulary which used terms such as luminosity and light. Accordingly, his *magnum opus* is called *Ḥikmat al-ishrāq*, 'The philosophy of illumination.' In a text he wrote later, *al-Mashāriʿ wa-l-Muṭāraḥāt*, 'The paths and havens,' the previously-mentioned symbols appear in the introduction. The passage can be translated as follows:

> When the student has fully grasped this way of thinking,[43] then let him commence with scintillating practices according to the judgment of the Custodian of Illumination, until he himself may see some of the principles of illumination so that the foundations of the matters become resolved for him. As for the three before-mentioned forms in 'The philosophy of illumination,' they are ***XYZ***. Understanding them is only granted after illumination.[44]

I used ***X, Y,*** and ***Z***, as placeholders for these symbols. The whole passage finds an equivalent in *Ḥikmat al-ishrāq*, which I will cite here too, to make the passage more understandable:

> I exhort you to preserve this book, to keep it safe and guard it from those unworthy of it. […] Give it only to whoever has fully grasped the method of the Peripatetics, a lover of the light of God. After commencing, let him practice for forty days, abstaining from meat, taking little food, concentrating upon the contemplation of the light of God, most mighty

[43] Suhrawardī sets up a difference between the philosophy of everybody else and his own. The former is considered Peripatetic and discursive, his own is illuminative and intuitive. For Suhrawardī they are not competing but different stages; one first needs to master Peripatetic philosophy before illuminative philosophy can be practiced. Cf. Suhrawardī, *The Philosophy of Illumination* [= Ḥikmat al-ishrāq], Translated by J. Walbridge and H. Ziai, Provo: Brigham Young University Press, 1999, pp. 3, 170 fn. 12.

[44] Suhrawardī, *al-Mashāriʿ*, in *Opera Metaphysica et Mystica* [= Oeuvres Philosophiques et Mystiques / Majmūʿa fī l-ḥikma al-ilāhiyya], Edited by H. Corbin, 4 vols., Orig. publ. 1945–1970., Tehran: Institut franco-iranien, 2009, vol. 1, pp. 194–195.

> and glorious, and according to what the Custodian of the Book commands him.[45]

These passages describe certain instructions for Suhrawardī's students about the circulation of his book *Ḥikmat al-ishrāq*. This book is not to be handed out until a person is already an advanced student of philosophy, with knowledge of books by, for example, Aristotle and Ibn Sīnā. Then begins a forty day trial period of asceticism and meditation. The final decision regarding the admission of a candidate to the next round is ultimately in the hands of a 'custodian' (*qayyim*); a term seemingly implying Suhrawardī nominated an heir to lead a group of initiated followers. In this context, Suhrawardī shares three symbols that are supposed to convey a key message about *Ḥikmat al-ishrāq*, and the knowledge of the symbols is only granted to the initiated.

In total, I collected digital versions of seven manuscripts and one printed edition of *al-Mashāriʿ wa-l-Muṭāraḥāt*, so that we may compare the symbols in different manuscripts. This step was done quite conventionally; by first collecting references to different manuscripts from the introduction to the edition and from other scholarly articles, and then setting out to acquire digital copies of them one by one. As a second step, I reduced this material to a folder with one image file per document, giving each image the name of the origin. From a quick comparison of these images, looking at them in rapid succession, it became clear that the symbols were cause of confusion among those copying the text. Two manuscripts did not have them at all: in Ayasofya 2571, the text simply continues without the symbols, as though nothing was supposed to be there. In the case of Milli 32785, an empty space is left where the symbols ought to have been. These omissions prove that the copyists deemed the symbols as something extraordinary from the text, something to be added later. Consider also the placement of symbols in other manuscripts: in Ayasofya 2570 and Topkapi 3377, the text runs equally continuous as in Ayasofya 2571, and the symbols are drawn in the margin. In Leiden Or 365, they are also in the margin, with a *ṣaḥḥ* ('correct') to indicate it should be considered part of the text. Even though they should be considered part of the text, their placement in the margin signifies a paratextual quality. This is even true when the symbols are in the text block. In Laleli 2552, they span no less than four lines and function like an inline graphic or illustration. Only in Arabca Yazma A 4302 are the symbols truly inline. However, when we look at the size and color, we see signs that even in Arabca Yazma's manuscript the symbols are given a special place,

[45] Suhrawardī, *The Philosophy of Illumination*, p. 162. Translation adapted.

as they are written similar to the chapter headings, being slightly bigger and in red. In fact, in all but Leiden's manuscript, the symbols are drawn in the ink of the rubrics (red for most but gold for the Topkapi one).

To analyze the six versions of the symbols I had, I loaded the relevant images onto my iPad. I then opened the application Vectornator and opted to begin a new drawing. Once a blank canvas was loaded, I inserted one of the images by clicking on the flower-like symbol in the menu on the top right. Once loaded, I locked the layer with the photo, so that I could not accidentally move or change it. Already having the photo like this was a vast improvement over looking at it on my computer, as I could zoom and rotate to my heart's content.

From here, the idea is fairly simple: trace the symbols from the photo using the pen tool to get the vector shapes of the symbols. This provides several benefits. First, by making vector shapes out of the symbols, you get a sharper version of the symbols, for which it is easier to see what exactly they are supposed to depict. In fact, I found that the very exercise of redrawing the symbols forces you to consider carefully what exactly the shape of the symbol is. For example, since the symbols are written fairly small in a few manuscripts, and since ink leaves a certain thickness on the paper, if the scribe makes two strokes very close to each other, the ink of both strokes might muddy. Given a photo with only so-so quality will exacerbate this effect. By zooming in and retracing the symbol, you might find that there are in fact two strokes, and to indicate this, you will likely make the gap between the strokes slightly bigger than is visible in the photo. Second, once you have vector shapes, you can freely enlarge them, move them around, rotate, mirror or distort them otherwise; you can try to make a bigger shape by composing them or try to make smaller shapes by deconstructing them. Since you can copy and paste the original vector shape over and over again, you can try out many different things with relative ease. The ability to give them different shape colors and turn the borders on/off is extremely useful, especially when deconstructing the shapes or overlaying different versions of the shapes.

To do the tracing, I follow several conventions. First, already at the drawing stage, I do some deconstruction by drawing the symbols into meaningful, small shapes. This is better, since it is very easy to combine different shapes into one or simply give them the same color to give the appearance of one uninterrupted shape, but it is more laborious to break shapes up. Second, I draw every shape on a separate layer and keep all layers for one symbol together. To distinguish between different manuscripts, I keep the photo layer and the symbol layers grouped together. Third, I try to draw the shapes as sleek as possible, since it is easier to expand them and have them retain the correct shape than it is to shrink them and maintain their shape.

Drawing a set of three symbols took me about twenty to twenty-five minutes. To do this for five manuscripts and a printed edition is, then, a non-negligible investment of time, but it can easily be done after office hours as it is a fairly relaxing exercise. I exported the results to SVG and loaded this into the Illustrator to prepare the illustrations below, which I saved as PDFs.[46] The first illustration gives an impression of the different versions of the symbols.

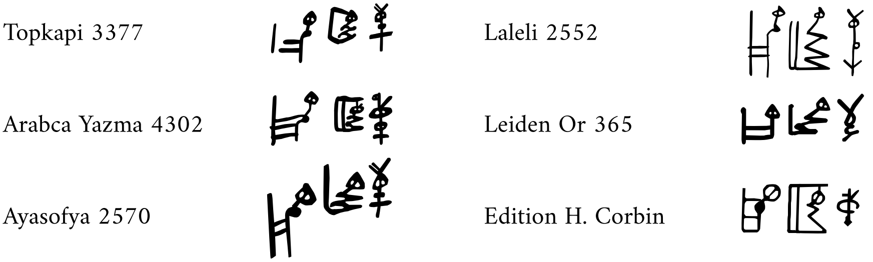


FIGURE 4.1 Suhrawardī's symbols redrawn from five manuscripts and an edition

What is instantly clear is that the shape of the symbols is not uniformly agreed upon, meaning we are faced with a double-layered puzzle. We need to find out the original shape of the symbols, and we want to figure out what those shapes could mean. Further, by comparing Corbin's version with the manuscript versions, we can recognize what Corbin did when he prepared his edition. The symbols in Corbin's edition are clearly based on Arabca Yazma's manuscript (or a manuscript associated with it) and go beyond it by idealizing the shapes. Corbin apparently saw them as geometric shapes and, therefore, emphasized this in his rendering, only making use of straight lines, squares, and circles.

We can break them down one by one. The following illustration gives the six versions of the left symbol with each element that is in common in the same color. The two letters underneath are an abbreviation for the source from which they came.

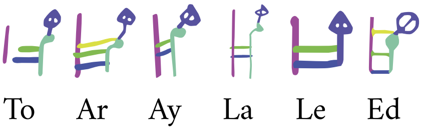


FIGURE 4.2 Deconstruction of the left symbol

[46] SVG stands for Scalable Vector Graphic. This image format is much used in web technology.

What is now apparent is that the symbol is seemingly made up of Arabic letters. On the left is an *alif*, common to all versions. Then there are two horizontal strokes like a *ba*. A third stroke is only present in the versions of Arabca Yazma and Corbin. A *mim*-shaped stroke appears on the right, although Leiden's manuscript apparently does not have it. On top, common to all, is a *ha*. The greatest common factor can be found in the manuscripts of Topkapi, Ayasofya, and Laleli. It may be immediately noted that those three versions are still quite different. Topkapi's left symbol has its *alif* detached, while Ayasofya is thick and slanted to the left, and Laleli is thin and long.

But the middle symbol shows a different story when we look into it.

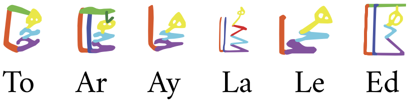


FIGURE 4.3
Deconstruction of the middle symbol

For this symbol, Topkapi, Ayasofya, and Laleli are not similar at all, while the manuscripts from Ayasofya and Leiden are almost identical. An *alif* shape (in orange), is common to all, and so are the two *ḥa* shapes and a *ha*. A *ba* shape on top is only visible in Topkapi, Arabca Yazma, and the edition. A second *alif* is only present in Arabca Yazma, Laleli, and the edition. Then, there is the strange case that Arabca Yazma has an additional v-shape at the top, maybe like an *ʿālamat al-ihmāl*, a sign to indicate an unpointed letter. Laleli has an additional *ḥa*-shape. The greatest common version of the symbol is, then, presented in Ayasofya and Leiden, with Ayasofya's version looking the most like the others.

For the symbol on the right, we have the following comparison:

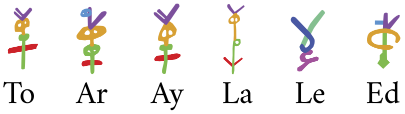


FIGURE 4.4
Deconstruction of the right symbol

For this one, Leiden's manuscript simply has something entirely different; a *lam*, *alif*, and *hamza* or *ʿayn*. Equally, Corbin's edition made a bit of a mess of this symbol. In fact, only using the other symbols was I able to discern the different elements and the correct shape of Corbin's version, and thus we may still identify several elements with elements from other versions. A central element is a *ha*, which is followed downwards by a *mim*. In Arabca Yazma, this *mim* looks more like another *ha* with a long tail, and in Corbin's edition it is merely a stroke with an additional blob at the bottom. A *ba* (in red) is common to many, though Laleli's one has an angle, almost like the v-shape we saw before. Such a v-shape, in fact, is present in all of them at the top. Something else is going on there too, about which there is little agreement. Topkapi's and Ayasofya's manuscript suggest an additional stroke, while in Arabca Yazma there seems to be a tiny *ha* of some sorts, which translates into Corbin's edition as a horizontal dash. It may again be noted that Ayasofya's manuscript has the most common depiction of the symbol.

Using this interpretative angle, I wish to single out the versions of Ayasofya and Leiden as particularly representative. Given that they show the most stable elements in their symbols, could it be that they show the symbols in their most correct form? If this is so for the left and middle symbol, how did Leiden get such a different symbol for the right one? I will leave these two questions open, but I will elaborate on the main question: what do the symbols mean? According to our work, they are constructed out of Arabic letters, and perhaps this is a factor for their meaning.[47]

If we follow Ayasofya's symbols, the sequence of Arabic letters would be the following: *ha*, *mīm*, *ba*-shape, *ba*-shape*, alif* - *ha*, *ḥa*-shape, *ḥa*-shape, *alif* - *v*-shape, *ha*, *mīm*, *ba*-shape. With "-shape," I mean that this letter is ambiguous since the base shape is shared by different letters. With one dot underneath the *ba*-shape, it is in fact a *ba*, but with one dot above it is a *nūn*, with two dots above it is a *ta*, with two dots underneath it is a *ya*, and with three dots above it becomes a *tha*. A *ḥa*-shape can also be a *kha* and *jīm*. If we follow Leiden's manuscript, the last symbol consists of *alif*, *lām*, *hamza* or *ʿayn*.

The use of letters to construct a symbol may be significant for an anagram or if they are used for their numerical value. If it were an anagram, I can only read أخوة باب الحماه in it: 'The brothers of the gate of Hama.' But as I indicate, this would suppose one more letter, the *waw*, and it would require the use of the Leiden version for the symbol on the right, whereas our analysis seems to suggest Ayasofya to be a better representative. Hama is a city in Syria, about 120 km south of Aleppo and about 45 km north of Homs. I do not know what its significance might be, other than a name for the initiated group of followers which Suhrawardī alludes to in his *al-Mashāriʿ* and *Ḥikmat al-ishrāq*. For numerical values using the Abjad system, and relating this to numerology, the options are too diverse, the results too speculative to make any formal

[47] There are other leads still open, too. Given Suhrawardī's interest in astrology and the occult, his symbols may have derived from that literature. Additionally, the style is reminiscent of cryptographic alphabets such as analyzed in Monteil, V. "La Cryptographie Chez Les Maures." pp. 1257–1264 in *Bulletin de l'Institut Français d'Afrique Noire* 13, no. 4 (1951).

statement. The bottom line is, however, that by changing specific characters of paleographic interest into vector shapes, we can find out more about them in an easier and visually more obvious manner. Once converted to SVG, it is easy to include them in an online publication, such as an edition or catalog. To do so, however, requires a lot more understanding of how text and graphics work together in the digital world. This will be the topic of the next chapter.

CHAPTER 5

Philology: Standards for Digital Editing

Among the most attractive aspects of digitized manuscripts is their use for philological purposes. When setting out to edit a text, you will invariably find that its manuscript copies are scattered over different libraries, and it is costly and impractical to visit them one after the other. Indeed, for most of the twentieth century, scholars already collected photographs or microfilms of their manuscript copies in order to lay them side by side and work on them without the pressure of opening hours or erratic viewing policies at libraries. In the digital age, bringing surrogates together in one digital environment is a godsend. All too often, discussions about digital critical editing are about its end result and the thorny question whether or not we should be making a 'digital edition.' Since that topic is worthy of a handbook in itself, we will not go into it in great detail. Rather, we shall look at our digital workflow in a more general way, paying attention to the standards we ought to use and how the photos of digitized manuscripts can fit in that workflow. We start with a general introduction to standards in the world of computers. Then we discuss how to move from simply typing text to building a critically marked-up text, and finish that section by discussing several typical use cases. Then we look at how images can be incorporated into the workflow. We finish with notes on how to preserve and continue to work on the materials we produce along the way, and finally revisit that delicate issue of digital editions.

A standard is nothing more than a detailed agreement to do something in a particular way. Abiding by this agreement is important because if large amounts of people do things in the same way, they can benefit from each other. Living on the other side of the Atlantic will make it rapidly clear just how important standards are—for length, weight, temperature, shoe size, where to place a comma in a number, how to display afternoon time … the list goes on.

People can easily spot if something is off, but a computer cannot. If I say the temperature is 20 degrees and it is summer, an American will understand I am talking in Celsius, not Fahrenheit. A computer controlling a heating system, however, will take that input without complaint, turning the system at full blast to combat the supposed freezing conditions, if it is set to Fahrenheit. If a person controls the input to the system, there is a good chance the person will convert the temperature to Fahrenheit before putting it in, but if it is only another computer that hands over the temperature input (say, the server of a

© CORNELIS VAN LIT, 2020 | DOI:10.1163/9789004400351_007

national meteorology institute), there will be no possibility at both ends of the communication to intelligently evaluate the message. And here is the catch: most of the times in our work, we talk to each other by having two computers talk to each other. If I use a messaging app, on a technical level, I am not directly communicating with someone else, but I am relaying keyboard inputs to my smartphone, which in turn talks to someone else's smartphone, which then relays differently colored pixels to that other person. I once tweeted the emoji for the Kaaba, the black box-shaped building in Mecca toward which Muslims pray. From people's responses, it became clear that they did not have this emoji installed, and their computer could, therefore, not display it as a little depiction of the Kaaba: 🕋, but instead displayed the standard sign for a character that is not available, which is a black square with question mark like this: �. Obviously, that is not a picture of the Kaaba, but in a moment of digital poetry people thought that this square with a question mark is an excellent symbol of it. You will probably have concluded by now that it is of the greatest importance that computers run their operations by strictly defined yet broadly accepted standards.

We will look at old standards and new standards, open standards and closed standards, standards defined by the tech industry, and standards defined by academic stakeholders. There will be standards that you consciously use daily, those that do their work under the hood of your computer, those that you may not need to bother with ever. We will even get a feeling for good and bad standards. In this chapter, we will go through virtually all parts of a typical workflow, proceeding from the smallest, most computer-related scale of bit and byte to the widest, most human-related scale of publish and cite.

1 File Formats

We can think of words, sounds, and images in the abstract, but we also know that when we want to embody them, they require a certain form. That form, in turn, dictates the way it is supposed to be used. If music is recorded on vinyl, you cannot push it into a CD player and expect to hear the music, and a manuscript is used differently from a scroll, in order to read the entire text. It is the same with files on a computer: they are in a certain format, which stipulates rules for a program as to how to open the file correctly. For example, if you tried to open a .jpg file with a simple text editor (like *Notepad* for Windows or *TextEdit* for macOS), you would not see the image visually, but you would see a seemingly random string of letters and other characters along multiple lines.

That is because the text editor expects the file it is opening to be a text, and it will convert whatever bits and bytes the file is made of into the text displayed on the screen.

In the world of computers, an important distinction is made between text file formats and binary file formats. For text, you may be thinking of a few notable file formats such as .txt, .doc (and .docx) and .pdf; however, of these examples, only .txt is a text file format, and the others are binary file formats. All text file formats admit to two restrictions: (1) they only contain plain text (written characters broadly understood) and (2) this text is organized in lines separated by a *new line* character. In a happy moment of agreement, all operating systems can open text files natively and opening any sort of text file in any sort of text file application will display the contents of the file. Text file formats are mostly defined to hold textual or numerical data (examples are .txt, .xml, .json, and .csv) or to contain code (examples are .py and .js). This is in stark contrast with binary files, of which each has a unique way of storing their data. There are binary file formats that are quite common, such as .jpg or .mp3, but many more file formats can only be written and read by a particular program. In fact, many of them are proprietary, and the particular way they encode data is a trade secret jealously guarded by the company which owns it. For example, .mellel is a closed, proprietary format used by the macOS word processor Mellel, meaning that only the owners of the file format know exactly how it works. Opening it in a plain text editor will result in a garbled mess.

The most important text file formats are .txt and .xml, on which more will be explained in the section about mark up. Microsoft Word's .doc is a binary file format. This is due to the particular way Microsoft has encoded the layout and markup possibilities for a Word document. For years, its encoding was not disclosed. Its successor, .docx, is a strange mix of text and binary: on its own, it is a binary file, but if you change the .docx extension to .zip and unzip it, a collection of files and folders become available that are mostly text files in the form of xml. The same cannot be said of Adobe's .pdf file format, a big favorite in the humanities for its function as a print-like digital file,[1] which is squarely a binary file format. When opened in a PDF viewer, everything looks nice and shiny, but under the hood, .pdf can be quite a mess. Therefore, even from PDFs with digital text (i.e., selectable and searchable), it is not always easy to extract the text automatically. Here, we stumble upon a practical difference between text files and binary files: text files are by far preferable when we want to do

[1] For as long as (and long after) the humanities are figuring out the transition to the digital world, especially as long as there is fear that publishing digital-only 'does not count', I suspect that PDF files will function as a hybrid: they are existentially born-digital but with the look and feel of a print material.

complicated or automated manipulations on the file, but even though binary files carry more bloat with them, they can present the user with a rich experience right out of the box. The .pdf format is a good example of this. By saving something in .pdf, you can be assured that when you share the file, the other person will see it in exactly the same way as you created it, and you will also be assured that the file cannot be (easily) altered. Compare that with sharing a .doc (or .docx) file. Perhaps you created the document with a particular font that the other person does not have on their computer. In that case, when that person opens the document, Microsoft Word will use a different font to render the text, and this may cause all the lines and paragraphs to shift slightly. If there is a complicated layout with tables, footnotes, and images, this may not line up anymore.

Text file formats are not entirely without markup. For example, spaces, punctuation, and indications of where a line ends (and a new one begins) are native to text file formats even though we might not strictly classify them as plain text. This is especially true for the *new line* marker, which can be inserted easily by hitting the *Enter* key on your keyboard. We often do not think of it as a character, yet it is one of the most frequently exploited characters in programming, as we can give a command to do something for every string of text we find on a separate line. *New line* also allows .csv files (comma separated value) to store a primitive table in which each line represents a row, and every column is separated by a comma. Codes, such as .py for Python or .js for JavaScript, are also text file formats, and some programming languages see the *new line* as an indicator for a new command to be executed.

Text file formats form a good mix of being human-readable and machine-readable. This is why they are excellent for storing code and textual and numeric data. Programmers use text file formats all the time. Meanwhile, binary files are better suited for presenting rich media and, therefore, they are the kind of file format that is more often encountered by the general public. We in the humanities will have to straddle both worlds. For example, we can get far by using digital photos of manuscripts in a conventional manner by simply loading them in an application that displays the image on our screen. If we want to go further, as we will do in Chapter Seven, we basically have to turn the image (a binary file) into a text format, by reducing the image to a few million triple values representing the mix of red, green, and blue for each pixel. Only then will we able to do computations and automated transformations on it.

Images are, of course, of special interest to us. They can come in a great variety of formats. The most fundamental digital data of photos is called a raw image file, and this can take a great variety of file formats, depending on the factory and the type of camera you are using. This cannot be used in and of itself for viewing and manipulating. Instead, from here, two file formats are often used: .tiff and .jpg. The former can store the image in lossless quality, meaning that no compression is applied that would reduce the quality of the raw image. The latter has variable compression, allowing for a trade-off between file size and quality, which is especially attractive if the images need to be available on the internet. In Chapter Four, I already introduced vector images as a different kind of image. Rather than storing an image pixel by pixel, such images store objects mathematically by begin point, end point, and the curve of the line in between (and from there, other aspects such as line style and fill). Since these images are defined rather than prescribed, their actual rendering on a screen is done on the fly, which ensures that the image is always sharp and crisp, no matter how far you zoom in. Obviously, this format is not suitable for photographs, but more so for digital drawings. An interesting aspect of vector images is that they can be stored in text file formats. One such format is .svg, which is often used on websites. Perhaps surprising, but .pdf is principally also a vector image file format, especially suited for print or print-like purposes. As we have seen, pdf is a binary format.

It is an essential skill to identify the right file format for the job at hand and, by extension, it will be very advantageous to be able to convert files from one format to another. We can distinguish four levels to do that: one is to use the functionality provided by the operating system itself, another is to use an application with a graphical user interface, the next is to use an application that can only be used with a command line in the terminal, and the final option is to program a script yourself.

For simple image manipulation, a myriad of applications can help you. I have found Resize Sense (macOS only) to be a useful application with a GUI to do batch rotating, resizing, and renaming, for example, after I have taken a few hundred photos of a manuscript. ScanTailor is a wonderful tool for page splitting, auto rotate and dewarping, and (if desired) turning photos into sharp black and white images.[2] Adobe Acrobat Pro can be used to extract all pages as individual photos from a PDF or create a PDF from individual photos. Occasionally, for example, if you receive photos in an unusual file format or if they have an unusual size, you might have to resort to the very powerful ImageMagick, which you can instruct to do things from the command line.

The same goes for taking data from the internet. You might just simply click a button that says 'download' or you might right-click and select *Save As*. You might also have to go dig deeper using your browser to find the page resources.

[2] This is especially useful for sources that are originally in black and white, such as printed publications. Returning them to black and white will often times make them more legible and smaller in file size. If there is good OCR software for the script the text is written in, this would be an excellent preparatory step.

If that is too tedious, perhaps because you are after a whole number of files, you could use a plugin for your browser with a GUI, or a stand-alone application called a download manager. Through all kinds of technical features, these applications can truly reduce the time to download something. They can also stop and resume downloads and make sure you are not banned for overusing, among other things. A very powerful but command line only software is *wget*. And lastly, there is the possibility to fine-tune it exactly as you want or need with web scraping tools and coding it yourself.

It is best practice to slowly escalate the kind of tool you use. Usually, this is dictated by the limitations you run into, and in this sense, you will notice soon enough when you need a different approach like a command line tool or a self-coded solution. Let us consider one example here. Say I shot eight photos with my iPhone, and I now want to make one PDF of them to share it with someone else. Those eight files are shot in .heic file format and amount to 9.1MB on disk. Once I have them on my laptop, I can convert them in several different ways, which I briefly summarize below, with the method on the left and the file size of the resultant PDF on the right:[3]

TABLE 5.1 Different results for converting an image

Native support of the operating system	
Right Mouse Click > Quick Actions > Create PDF	131,7MB
Opened in Preview in one window > Print > As PDF	131,7MB
Using Automator 'New PDF from Images'	130,1MB

Using an application with GUI	
Using iMazing HEIC Converter for conversion to .jpg, then Right Mouse Click > Quick Actions > Create PDF	39,2MB
Using Adobe Lightroom for conversion to .jpg, then Adobe Acrobat for conversion to .pdf	25,1MB
Using FreeToolOnline-website	919KB

[3] The data produced here is, of course, heavily dependent on the exact version of hardware and software. I did this test with the following equipment: iPhone 7 Model A1778, MacBook Pro Retina 13" Mid 2014. Software: macOS Mojave 10.14.1, iMazing 1.0.7, Adobe Photoshop Lightroom CC 2.0.2, Adobe Acrobat Pro DC 2019.008.20071, http://freetoolonline.com/heic-to-pdf.html accessed on 11-11-2018, ImageMagick 7.0.8-14.

TABLE 5.1 Different results for converting an image (*cont.*)

Using a command line tool	
Using ImageMagick for conversion to .pdf	295,1MB
Using ImageMagick conversion to .jpg, then to PDF	16,3MB

On macOS, the most direct way, especially if you are only concerned with a quick result, is to right-click on the photos and select from the dropdown menu the submenu *Quick Actions* and then click on the option *Create PDF*. This will create a PDF file in the same folder as the photos are and will put the focus on the file name so that you can immediately type out a different file name. Though it is a convenient way, you end up with a file that is extraordinarily large. You could also open all the photos in Preview, the standard image viewer of macOS, drag their thumbnails all into one window, and then simply print as PDF, but this results in exactly the same file size, this time with some white space on some edges. A slightly more advanced approach but still using the functionality of the operating system only is to use Automator and its function *New PDF from Images*. This gives, a bit surprisingly, a slightly (but not much) smaller file size.

Clearly, only in the direst of situations and only if you need to create one PDF would this be an acceptable solution,[4] so the next step is to use a more specialized program with a point-and-click graphical user interface. The free application iMazing HEIC Converter helped me out here. It allows a quick drag-and-drop of .heic files and converts them to .jpg. Of course, that does not make us a PDF, so we still need to use the function native to macOS to *Create PDF*. iMazing manages to convert the HEIC files to JPG, but it does so at the cost of increasing the file size fourfold. The upside of this workflow is that once they are in JPG format, the *Create PDF* function adds no extra file size to them. iMazing is on the easy and cheap end of the spectrum of consumer software, so let us see how software on the other end of the spectrum fares. The software suite of Adobe is the obvious candidate, whose sticker price is beyond most mortals' reach, but most universities will likely offer it for free to students and employees. The application for creating PDFs, Acrobat, does not support HEIC files, so we will first need to use Lightroom to convert the files to JPG.

[4] One way of sharing such a file is to place it in your Dropbox (or equivalent) account and share a link.

Combining both applications will result in a PDF that is significantly smaller than the previous workflows, all the while retaining quality.[5]

An entirely different approach is to use an online resource to do the conversion for you. Though a few websites advertised support for HEIC files, they actually did not. In the end, 'FreeToolOnline' did work: I only had to visit the website, upload the images, and click 'convert.' This gave a shockingly small file size, small enough in fact to attach to an email, but needless to say, this was at the cost of reduced image quality. There are no options to control this, and the reliance on a website is itself cumbersome.

A final resort is the command line tool ImageMagick. However, its immediate conversion to PDF did not fare well: not only did it result in an absurd file size, but the colors were completely off. Converting to PDF through a middle conversion to JPG gave a very satisfactory result, with the smallest file size and no compromise in resolution.[6] It should be noted that in fact, all the methods described here resulted in slightly different colorations, which will be a high priority issue for curators but is less of a concern for our private usage. For example, when doing a visual comparison, ImageMagick's JPGs were slightly lighter than the original HEICs, while Lightroom's JPGs were slightly darker than the originals. We see, then, that the different processes we can apply to convert file formats will lead to different results. Being aware of the possibilities, then, is very useful.

Let us do a similar comparison for text. In this example, I started with a Word document, containing an introduction and edition of a short 17th-century epistle. The document contains English, Arabic, and Persian, with a combination of footnotes and endnotes.

TABLE 5.2 Different results for converting a text

Method	**File size**	**Correct characters**	**Foot-/Endnotes**
Original .docx file	177kb	yes	yes
Using Microsoft Word, saving as .txt	79kb	slight mixup	only text
Using Pandoc	81kb	yes	yes
Using Python library 'python-docx'	66kb	slight mixup	no

[5] Lightroom changed the PPI from 72 to 240. This obviously does not come with quality increase, but does increase the file size and gives a false impression of higher quality if one were to 'fly by instruments.'

[6] I first used this command: magick convert *.HEIC -set filename:f '%t' out/'%[filename:f].jpg' Then this: magick convert *.jpg manuscript.pdf

If we simply open the file in Word and click File > Save As …, we can save the file within Word to a .txt format. The line breaks are preserved, and so are the indentations at the beginning of every heading and paragraph, but no further distinction for headers or any other formatting is made. There is a slight mix-up when the Arabic and Latin characters are mixed, but more importantly, while the text of the notes has been preserved, there is no reference to where they belong in the text.

Pandoc is a powerful command line tool to transfer different text formats.[7] It also supports TEI-XML, a standard we will soon discuss. With this .docx file, it creates a .txt file that is remarkably useful. Virtually all formatting is preserved: for example, headings are distinguished by a following line with either = or - signs, italic text is indicated with a * sign, and the notes have a referent in the text. For footnotes, Pandoc uses [^x], with X the number of the footnote. Pandoc places this both at the position in the text and just in front of the note (at the end of the file). For endnotes, Pandoc uses [x^], so that we can quickly tell it apart from footnotes. Such formatting is useful since we can then manipulate it with a programming language and perhaps later even restore it to a .docx format.

Python is a programming language, and python-docx is a library you can download and run from within Python (more on Python in Chapter Seven). In this example, I used the most basic command to extract all text, and apparently this does not preserve formatting, not even the text of the notes. In this example, then, it does not particularly shine, but unlike the other options, it has tools to access the specific parts of the document: for example, if we only wanted to extract the English introduction, we could have written instructions for Python to do just that.

2 Encoding of Text

We previously established that text, broadly defined as human-readable characters in a wide sense, plays a particular role in the digital world, treated strictly differently from all other data. But if the computer can only handle zeros and ones, how is the text itself defined? Let us discuss this from the ground up.

A bit is the smallest unit stored on a device, and it can merely store as 'yes' or 'no,' 'on' or 'off,' 'zero' or 'one'. The next smallest unit is a group of eight bits called a byte. Since each bit inside this group can be 0 or 1, there are 256 (2^8) combinations. There is no need to make each combination represent a number

[7] I simply used the command: pandoc -f docx -o pandoc.txt Original.docx

because the numbers zero to nine are sufficient to express any number in our decimal system; with multiple bytes, we can encode as large a number as we like. The other 246 slots are, therefore, available for other things such as the alphabet and punctuation. Imagine these combinations of zeros and ones written out underneath each other, with 00000000 at the beginning and 11111111 at the end. If such a list has a character next to it, we have a simple way for a computer to look up what to print to the screen when encountering a specific combination of zeros and ones. An early lookup table that gained tremendous popularity and is, therefore, still shaping modern encoding is ASCII (American Standard Code for Information Interchange). ASCII only defines a list of 128 elements; in other words, it can suffice with 7 bits.[8] In ASCII, 1000001 is 'A' and 1100001 is 'a.' ASCII has some basic punctuation and extra characters, totaling 95. If you have an American keyboard, you will notice that all of them have survived till this day: 47 double-labeled keys plus the spacebar. The other combinations of zeros and ones define the so-called control characters. Some of them can also directly be typed, such as Tab, Return, Escape, and Backspace. ASCII defines a sorely insufficient list, as characters such as é, ç, ß or ø are not in it and, therefore, cannot be encoded despite their prevalence in European languages. Scripts other than Latin are simply not available. It is, however, still good to know of this heritage, since there is a small chance you will encounter ancient software that only recognizes ASCII. By giving it differently encoded text, in other words, that uses a different lookup table, the software will output something completely wrong. In my own work with Arabic, this often happens in a non-obvious way, where the Arabic is displayed correctly in the PDF but once selected, copied, and pasted, it shows garbled characters.[9]

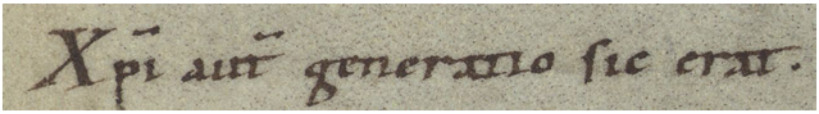
The one encoding you need to know of is UTF-8, short for 8 bit Unicode Transformation Format, often simply referred to as Unicode.[10] Most companies and organizations in the tech industry support or recommend Unicode, which has made it the de facto standard. By using multiple sets of eight bits, UTF-8 can define up to 1,114,112 characters, of which so far 137,374 have been assigned. A thorough discussion of Unicode is outside the scope of this book, but it is worth noting that more and more ancient scripts are supported in Unicode. This is a major boon for us, as we can digitally transcribe manuscript sources in a digitally-native way without any hacks or tricks. Nonetheless, it

[8] This is a remnant of a computer era in which the standard of 8 bits as a byte was not fixed yet.

[9] A technical term for this phenomenon is mojibake, a term with Japanese origin.

[10] In the rare occasion that a program is not reading your text file correctly, you might want to look into UTF-16 and UTF-32.

remains good to acquaint yourself with the Unicode support of your particular script, to understand its history, benefits, and drawbacks. As an example, take the following excerpt:[11]



Xpĩ aut̃ generatio ſic erat.

FIGURE 5.1 Encoding a manuscript

This represents five words from the Gospel according to Matthew, Chapter 1, Verse 18. We could simply transcribe it as "Christi autem generatio sic erat." If we want to be more faithful to orthography, we could also transcribe this as "Xpĩ aut̃ generatio ſic erat," where we use one Unicode character for the ĩ, namely "Latin small letter i with tilde" (U+0129) and two Unicode characters for t̃, namely "Latin Small Letter T" (U+0074) and "Combining Tilde" (U+0303).[12] Computers that decode Unicode text will understand 'combining characters' as a character that should not receive its own space but should be placed above or below the previous character. Visually on screen, the result is one character, t̃. We further changed the s in 'sic' for ſ, which is the "Latin small letter long s" (U+017F). You may have noticed the transcription of 'p' as the second letter of the first word. This word was simply understood as consisting of the Latin letters X, p, and i, forming a *nomen sacrum*. If we, for one reason or another, wish to note its Greek origin as Chi, Rho, Iota, we could encode it like this: Χρι. Notice the use of "Greek Capital Letter Chi" (U+03A7) and not "Latin Capital X" (U+0058), even though visually they are indistinguishable. We can see the flexibility of Unicode: without installing any additional software, I can easily type out these different transcriptions, and I am assured that they are understood by most computers.

However, there are caveats. First, "Χρĩ aut̃" may at first sight appear to be a very accurate digital rendering of the manuscript evidence, but the digital

[11] "Reichenau Gospels." MS W. 7, Walters Art Museum, Baltimore, 11th c., f. 17r. The digital image is 5137 x 5948, showing one page at about 91,7MB. The cut is good and color balance excellent.

[12] This notation, capital U, the plus sign, and a combination of numbers and letters, is the standard way to refer to a specific Unicode character. This combination can be used to directly write the character. On Mac: Install the Unicode Hex Input keyboard, select it, and hold down *Option* while typing the combination. Windows: Hold down *Alt*, press + followed by the combination. Linux: Press *Ctrl+Shift+u*, followed by the combination.

tilde first goes up and then goes down, whereas the macron in the manuscript first goes down and then goes up, and there is nothing one can do against this. Second, even though the different ways of encoding are easy for people to understand, it can cause issues when a computer performs a search. If I transcribe the text as Xrĩ aut̃, a software that is not programmed flexibly enough might read the second word as being of four characters in length. It would then give a wrong character count. Or what if we search for “Xρι” and the software sees ‘ι’ as something different than ‘ι̃’? After all, U+03B9 is only one character while U+03B9 combined with U+0303 is two. Such cases show the importance of normalization: when a computer needs to manipulate the text, it should read ‘i,’ ‘ĩ,’ and ‘ĩ’ as the same thing. This kind of problem can occur more often when you work with non-Latin scripts. It becomes more frequent when the script becomes more obscure. This is because the available software does not significantly take into account such scripts. Here is an opportunity for us to take responsibility ourselves and work out issues as best as we can on our own. To do this, we need to operate on a level that allows enough flexibility—and by this, I mean we need to, if necessary, be able to parse text to simple text file formats such as .txt or .xml and be able to write code that takes the text apart in whatever way necessary.

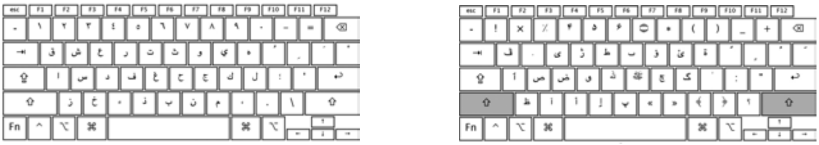
There is one more thing to discuss with regard to text encoding, and that is fonts. Unicode’s lookup table, which consists of more than 100k characters, is still an abstract list. A computer can only display it on your screen by using a font. A font is again a lookup table, but this time with Unicode on the left and a glyph on the right. Since fonts are restricted to 65,535 characters, no font can contain all of Unicode’s characters.[13] It is possible, then, that a computer first looks up which Unicode character belongs to the binary encoding, then looks up which glyph belongs to that character, and finds nothing. An empty box (sometimes described as a slab of tofu) will appear, or a black square with a question mark. This can be an issue when sharing documents. For example, when you write something in Word, the computer will ‘remember’ not only which characters you have typed, but also in which font. Somebody else will open that document on another computer, and the document will instruct the computer to render the characters in that font. If the font is not installed on the other computer, it will switch to a different font. If that different font does not contain the glyphs needed to render the encoded text, you are going to have a problem. This problem does not exist when exporting to PDF since you

[13] OpenType and TrueType, the most common font formats, do allow for one file to combine different font files, thereby being able to go over this restriction.

can instruct in the settings that all fonts used are to be included—in other words, the font you use is included in the PDF you make, which allows anyone else to see the document as intended.

There are two technologies that can make your life easier when it comes to entering digital characters. The first is to create your own keyboard layout. Consider your keyboard: every key has got something printed on it, like a letter Q or the number 5 and a % symbol, or the command Caps Lock. In the US, the standard keyboard layout is QWERTY, while in France they have AZERTY keyboards. A Norwegian keyboard has the letter Ø and Æ printed on the keys right next to L. In short, we know that it is a matter of choice that a Q or Z appears when we hit the key on the top-left next to Tab, or, in other words, that either U+0071 or U+007A is transmitted to the computer. We can, therefore, just as well create our own layout, where we stipulate exactly what Unicode character (or even combination of characters) is fed to the computer when we press that key. For Arabic, this has been an essential technology to speed up my typing, since the standard layouts for Arabic have the characters under the keys that have no relation to the letters on a standard US keyboard. My starting principles were to be able to type Arabic as though I was transliterating using the standard QWERTY layout and to have a layout that has everything on board so that I do not have to move my hands away from the keyboard and use the mouse to click through menus. Variations on a letter, such as emphatic letters, letters with hamza, letters particular to Persian, and letters without *iʿjām* are accessed by pressing Shift plus the closest Latin letter. A full set of vowels is included, as are a few ways to mark the beginning and end of a quote or a Koran verse. I even included an option to type something in Latin alphabet by holding down *Option* (on macOS). To make this, I used the application *Ukelele*, which has an intuitive GUI. This allows me to type Arabic smoothly system-wide no matter the application I am using.

After creating a custom keyboard layout for Arabic, I turned my attention to the other major challenge involved when writing about Islamic history: to transliterate Arabic in Latin. I used a slightly more advanced feature of Ukelele that allows you to enter a modus with an alternative keyboard layout only



FIGURES 5.2A AND 5.2B Keyboard layout for Arabic without and with Shift pressed

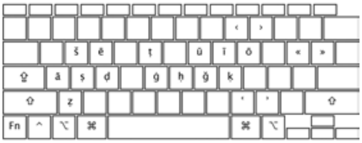


FIGURE 5.3
Keyboard layout for Arabic transliteration

when you press a specific key combination. I programmed three such combinations: *Option* + ., *Option* + *Tab*, and *Option* + \. The benefit of setting up three combinations is that if one combination is taken by an application, you can still use another one. This layout has replaced my normal layout, enabling me to type letters such as č, ē, ō, or ḍ.

Keyboard layouts are great for systemwide access to often-used characters. But if you need a rare character often for a particular project, or full names or titles that you keep using again and again, another highly practical technology that you can make your own is to run a text expansion application. This is a small program that runs in the background, where you can configure what you want to appear when you type only certain character combinations. This can be quite advanced with dynamic elements such as dates and markup such as italics, bold, or new lines. I have found it useful to set capital initials of names and titles to expand into their entirety. For example, "IA" is set to expand into Ibn ʿArabī and "NDT" expands into Naṣīr al-Dīn Ṭūsī. Since this happens at the very moment you finish typing the last initial, it feels natural, and you keep typing, only now you progress much faster since you do not have to type the unusual transliteration characters one by one.

3 Markup of Text

The previous example, where we considered a manuscript witness of "Christi autem," naturally raises the question: how can we avoid having to choose to encode either the textual, orthographic, or paleographic content and instead encode them together? For example, a user might be searching for "Christ," and we would want "Xρῖ" to show up as a hit. This, in turn, raises another question: how can we add other remarks, such as about grammar, semantics, topics, persons, and dates? Connecting transcriptions of different manuscript copies of the same text, noting variants of the same word or passage, is yet another way to ask this same question. The ability to register all these aspects digitally depends highly on your workflow.

If you work within a word processor, the possibilities are limited.[14] You now basically work straight from manuscript to published form, which can take not much more than the form of a printed critical edition in the style we are used to: a body of a text with a critical apparatus as footnotes. It is my impression that we in the humanities handle our standard word processor too often and too quick. Such word processors are alright for styling text documents and making them ready for basic printing needs,[15] but they are especially unsuitable for two common cases: note-taking and manuscript editing. Both these activities ideally operate on a level before publishing. To understand this better, we need to define the different stages of a digital workflow.

Andrews and Robinson, scholars with experience in digital editing, suggest the following stages:[16] Transcription → Collation → Analysis → Edition → Publication.

This differentiation is a good first step. The power of it lies in revealing the nature of the different parts of our work: with transcription, we work closely with the source; when we collate and analyze, we work with the digital encoding; and when we edit and publish, we work toward molding that encoding in an intelligent shape or form. Their separation of collation from the analysis is likely because of their preference for computational collation using software borrowed from Biology that supports phylogenetic analysis.[17] Their separation of edition and publication can be explained out of their preference for digital editions, from which they can suggest that one edition can spawn different publications (for example, one for the general public and one for scholars).

Apollon, Bélisle and Régnier, the editors of *Digital Critical Editions*, propose the following workflow:[18] Content gathering and preprocessing (including collation) → encoding for internal representations → transformation through algorithms → output

[14] Such as Microsoft Word for Windows, Pages for macOS, or LibreOffice for Linux.

[15] I am speaking here of private printing. Publishers might be willing to receive a Word file, but do not use them to do a print run for a publication.

[16] Andrews, T. "Digital Techniques for Critical Edition." pp. 175–195 in *Armenian Philology in the Modern Era: From Manuscript to Digital Text*, edited by V. Calzolari and M.E. Stone. Leiden: Brill, 2014, p. 176. Cf. Andrews, T. "The Third Way: Philology and Critical Edition in the Digital Age." pp. 61–76 in *Variants* 10 (2013).

[17] For example, take a look at Dekker, R.H., D. van Hulle, G. Middell, V. Neyt, and J.J. van Zundert. "Computer-Supported Collation of Modern Manuscripts: CollateX and the Beckett Digital Manuscript Project." pp. 452–470 in *Literary and Linguistic Computing* 30, no. 3 (2015).

[18] Apollon, D., C. Bélisle, and P. Régnier. "Introduction: As Texts Become Digital." pp. 1–34 in *Digital Critical Editions*, Urbana: University of Illinois Press, 2014, p. 4.

They seem to base their workflow on a methodology where you first establish a stemma before you even begin transcribing.[19] They further make the last phases sound a bit enigmatic, saying nothing about what these algorithms are supposed to be and refraining from the word 'edition' and 'publication.' This emphasizes that we do not need to get stuck on creating editions, but that we can use the encoded text for whatever purpose we see fit (such as a distant reading analysis). Compared to Andrews and Robinson, we can see an important conceptual difference: Andrews and Robinson want us to come at the sources without prejudice as to encode the manuscript evidence as raw as possible. In contrast, Apollon et al. want us to formulate our goals beforehand, so that we may include the aspects relevant to those goals directly in the encoding when we inspect and transcribe the manuscript evidence.

Lastly, we make a note of the model of Rehbein and Fritze, derived from their workshop on digital editing:[20] Data modeling → Transcription and Encoding → Publishing.

The advantage of this workflow is that it does not derive from the classical method of editing but considers the workflow anew from a digital point of view, most notable from the introduction of the term 'modeling.' As the two explain, this first step can be seen as a way to outline the entire project, as the model does not only follow out of the source material but also from the desired final product. For Rehbein and Fritze, this product is clear: a critical edition. How to move from encoding to publishing is left undiscussed, perhaps suggesting that the encoded text itself is an edition.

From their workflows, combined with the arguments made in this book, I propose the following synthesis:

> Digitizing → Transcribing ⇄ Analyzing → Publishing.

With this workflow, I want to foreground the photographic aspect of our work. The previous workflows focus heavily on the text as an abstract entity, thereby building in a blind spot for the material aspect of our work, and the photographs with which we work as part of that materiality. The digitization itself is best left to libraries and museums, but in our handling of these files, we ought to keep in mind the conceptual framework that I laid out in the first few chapters.

[19] There are cases in which this is a necessary decision, for example, when the manuscript evidence consists of hundreds of witnesses you need a stemma to find out which manuscripts are most useful to include in your edition.

[20] Rehbein, M., and Chr. Fritze. "Hands-On Teaching Digital Humanities." pp. 47–78 in *Digital Humanities Pedagogy: Practices, Principles and Politics*. Cambridge: Open Book Publishers, 2012.

I also want to bring out the interdependency of transcription and analysis, in which, I maintain, small samples of transcription ('test drills') can inform analysis such as coming to a good stemma and finding interesting aspects of the manuscripts and the text they contain. These findings can, in turn, influence the transcription and the decision regarding those aspects that require tagging during transcription. This workflow, it should be understood, also supports analysis of the photos and not just of the digital text extracted from them.

In agreement with Apollon et al., I see the step toward sharing our findings as a separate, final step. In this stage, we do not alter the encoded data we have but prepare a polished, publishable digital document based on the data and on the analysis thereof. It should be noted, however, that the data itself is also an outcome of the project and should be seen as a deliverable. This way, the workflow has a circular aspect to it: the end result of one project can be the grazing ground of the next.

Where would a word processor like Microsoft Word fit in? Working in Word would mean that we use it for basically all steps beyond digitization. Obviously, Word is not meant to carry such a burden, and as such, most of the analytical and editorial work only goes on in our head, and Word becomes a crude sketch pad out of which we slowly and dully draw our final product, which must be a printed edition since a word processor does not support much beyond that. Importantly, for everything that we use Word, we are irreparably altering the data, making it difficult to use the encoded transcription for anything else. For small projects that are solely geared toward a printed critical edition and deal with a text that almost certainly will not be of use for other bigger projects, a word processor may be the right tool. For anything more, we would do well to harness the digital power that a computer can offer when we serve it a digital and tagged text. This means that we should strive, if only in theory, to disentangle the digital data we make and save on our own computers and the digital documents (among them most importantly actual publications) we share with the rest of the world.

Perhaps on the other side of the spectrum stands the use of a graph database in which each word is a node, connected with another word that 'comes after' as an edge, also connecting to words in other manuscripts either by 'same,' 'equivalent,' 'variant,' or 'missing' edges, having attributes such as 'personal name' and 'abbreviation.' Data held together like this becomes impossible for us to read or even have a grasp on, but computers will glide through it in ways and at speeds impossible to achieve for human beings if they had a critical edition of the text under investigation.

Most people have settled on a middle ground solution, using plain text and adding special symbols that we can tell a computer to interpret in a certain way. Let us go over three different ways of doing this.

One simple way is to separate transcription from markup, as is done in TextFabric, a piece of software originally designed to analyze the Hebrew Bible but increasingly used to analyze other ancient corpora, such as cuneiform tablets. TextFabric uses no more than a collection of .txt files where a new line indicates a new bit of the text. So far, TextFabric has been used on the word level, meaning that the transcription is saved with every word on a new line, after which new text files can be created with an equal number of lines and information on the same line number as the word it informs. Thus, the simple fact of corresponding line numbers connects the markup with the transcription. For example, a file can be created on every line with the grammatical function of the corresponding word in the transcript. Once those files are created, TextFabric allows for advanced text analysis. As such, this approach is especially suited if you want to analyze a text deeply and can afford to invest time in preparing it and its markup in such a way.

Another approach is to have very light markup right there within the transcription. For example, the OpenITI initiative has developed mARkdown to structure an ever-expanding corpus of digital texts of the Islamic literary heritage, which runs over a billion words with even individual works comprising of several million words.[21] With text files that heavy, a very lightweight way of tagging is necessary. For example, one such text is a biographical dictionary with over 30,000 biographies.[22] If we need every biography tagged, then the length of a tag matters a great deal since it will be added 30,000 times to the file. This is what mARkdown proposes to do. It allows tagging headings, paragraphs and page numbers, poetry verses, dictionary elements, biography elements, names, years, and a few more other things. The tags usually start with three hashes. For example, the start of a biography of a man is indicated by ### $ and a woman is indicated by including ### $$. The end is simply inferred from the occurrence of a tag indicating the beginning of a new entry. Because mARkdown

[21] The corpus consists mostly of texts that were encoded by the *al-Maktaba al-shamela* initiative, which does not employ OCR but relies on two volunteers to type out the text (so-called double keying) with an automated comparison of the two encoded texts to weed out typos. In the name of mARkdown, the A and R are capitalized to note its use for Arabic. The name itself is derived from 'Markdown', a popular lightweight syntax to give plain text some formatting, which is used by the tech industry at large, worldwide.

[22] Romanov, M.G. "Observations of a Medieval Quantitative Historian?" pp. 462–496 in *Der Islam* 94, no. 2 (2017).

is project-specific, it also includes impurities such as tags that do not support research goals but are included to avoid malfunctioning of the particular technology the project is using.[23]

A middle ground between heavy and light markup is the third way we will look at, and this is by far the most popular one within the humanities. This third way is to use XML, which stands for eXtensible Markup Language. Plain text files that adhere to the rules of XML can be saved as a .xml-file. We will look at XML more closely in the next chapter. Let us here suffice with the two leading principles of XML. (1) An .xml-file consists of elements. Elements are easy to identify since they have a tag to mark the beginning and one for the end.[24] The beginning-tag looks like <N> with N a word of your choice, and the end-tag looks like </N> with N being the same word as before. (2) In a .xml-file, all elements are properly closed, and none overlap. If you consider these two principles, you may notice that it is not saying much. It is only declaring some sort of meta-structure, but it says nothing about how to tag a name or an abbreviation. The choice is left to yourself. Fortunately, these issues have been thought through before and the most significant result of that has been achieved by the TEI, short for Text Encoding Initiative. The "TEI Guidelines for Electronic Text Encoding and Interchange" (or simply: TEI) have been developed specifically for humanities research and give recommendations for how to call your tags and what and how to use it. Given the breadth of humanities research, TEI has ballooned to a size that can be, at first, overwhelming. However, projects seldom need more than a small subset of TEI, which makes it easier to familiarize yourself with the rules and tags that TEI offers. Even so, given that the same phenomenon can be encoded in multiple ways with TEI, the guideline itself requires a guideline to its best practices, preferably in the form of a workshop by a veteran TEI encoder.

On the following page, we see our example text encoded in TEI.

[23] Says Romanov: "While EditPad Pro [the particular technology the project relies on, CvL] handles large files very well, it has problems with long paragraphs (or, more correctly, lines). For this reason, long paragraphs are split into shorter lines, where each line starts with ~~ (two tildas [sic])."

[24] There is also an exception where beginning and end tags are combined into one.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>One Line from the New Testament, a Digital
Encoding</title>
        <principal>L.W. Cornelis van Lit o.p.</principal>
      </titleStmt>
      <extent>less than one Bible verse</extent>
      <publicationStmt>
        <authority>L.W. Cornelis van Lit o.p.</authority>
        <date>2020</date>
        <idno type="DOI">10.1163/9789004400351_007</idno>
        <availability>
          <licence target="https://creativecommons.org/
licenses/by/4.0/">
            <p>CC BY 4.0 applies.</p>
          </licence>
        </availability>
      </publicationStmt>
      <sourceDesc>
        <msDesc>
          <msIdentifier>
            <settlement>Baltimore, MD</settlement>
            <institution>The Walters Art Museum</
institution>
            <idno>W.7</idno>
          </msIdentifier>
          <msContents>
            <p>This line comes from the Gospel according to
Matthew, Chapter 1, Verse 18.</p>
          </msContents>
          <physDesc source="http://thedigitalwalters.org/
Data/WaltersManuscripts/ManuscriptDescriptions/W7_tei.xml">
            <p>This encoding was based on a digital
manuscript. The highest quality available gave an image of
5137 × 5948 pixels at 91.7MB, showing one page, the color
balance is excellent as is the cut. See @source for link
to original TEI file with codicological information of the
physical manuscript.</p>
          </physDesc>
          <history>
            <origin>
              <p>Probably from Reichenau Abbey, Germany,
11th century.</p>
            </origin>
          </history>
          <additional>
            <surrogates>
              <bibl facs=“https://iiif.archivelab.org/iiif/
W7000037600/514,3353,1270,166/full/0/default.jpg”>
                Image of only this line.
              </bibl>
            </surrogates>
          </additional>
        </msDesc>
      </sourceDesc>
    </fileDesc>
    <encodingDesc>
      <projectDesc>
        <p>This encoding is an example for the book ‘Among
Digitized Manuscripts’ published by Brill, 2020.</p>
      </projectDesc>
    </encodingDesc>
  </teiHeader>
  <text>
    <body>
      <p>
        <pb n=“17r” />
        <lb n=“15” />
        <name type=“person”>
          <choice>
            <orig>
              <choice>
                <abbr>
                  <foreign xml:lang=“grc”>Χρῖ </foreign>
                </abbr>
                <expan>
                  <foreign xml:lang=“grc”>Χρι</foreign>sti
                </expan>
              </choice>
            </orig>
            <reg>Christi </reg>
          </choice>
        </name>
        <choice>
          <abbr>aut </abbr>
          <expan>autem </expan>
        </choice>
        generatio
        <choice>
          <orig>ſic </orig>
          <reg>sic </reg>
        </choice>
        erat.
      </p>
    </body>
  </text>
</TEI>
```

The shape and color of this example are typical of XML: every tag has its own level, and this makes triangle shapes when there are tags within it ('children'). Tags, attributes, attribute values, and tag values all have different colors.

What may be surprising is how much space is allocated to things not directly related to the marking up of the text. Whereas in TextFabric the markup and text were separated, and whereas in mARkdown the markup could be said to be within the text, TEI allows us such an elaborate markup that it is starting to look like the other way around: the text is within the markup. When you look closely, you will notice that the document is made up of a teiHeader tag and a text tag. This is actually one of the strengths of TEI, in that it leads to a self-contained file. Given the ability of digital documents to be copied and transmitted freely, the context of this file might get lost. With the information in the <teiHeader>, someone who stumbles upon it will be able to figure out what it is, and what they can do with it.

The header starts with a file description, which, in turn, starts with a title statement, meaning the title of this digital document. In between the title statement and publication statement, I included an extent tag. A future reader, I can imagine, would open this file and worry that it is missing most of the text since it contains only a single line of marked-up text. With this tag, I can reassure those readers that that is correct. The publication statement then gives legal information, after which a source description takes up the remainder of the file description. In this source description, I identify the manuscript. Since we are working from a digital photo, I make this clear in the following physical description. Admittedly, my use is stretching the definition of the physDesc tag, since the TEI guidelines would actually want you to consider and describe the physical manuscript. The guidelines were written, then, with the use case in mind, where you have direct access to the manuscripts (or the authors simply think of digitized manuscripts as neutral windows onto the physical artifact, as described and argued against in Chapter Two). TEI only allows you to define digital surrogates in a separate tag, for which I have used an IIIF reference. This is again not exactly how it is supposed to be used, but currently, there is no good alternative. We shall discuss IIIF later on in this chapter. After this long description of the file, I included a snippet about encoding to bring to attention that this document is an example belonging to this book, and to explain the context of creation and the purpose of this file.

When we finally come to the text, I wrapped everything in a body and a paragraph tag as is customary. I included a page and line number to signal the larger context of the sentence encoded here. The first word is immediately the most complicated to encode. Certainly, there are other ways of doing it, and perhaps one way is preferable over the other for reasons of performance, such as the time it takes for a computer to parse search queries. In our case, I opted in the first place to mark up the word as a personal name. This way, if the document is considerably larger, we could do a query on the personal names to find out, for example, how many times a name is used or how many unique names are used. You could install a more complicated system, including an ID that links to an entry in a list that gives specific information about that person—something I did not deem necessary for this example. Second, I encoded two choices: one form of the word with Greek characters, and another, normalized form of simple Latin characters. Further, for the first form, I provided two choices again: either an abbreviated or unabbreviated form. Lastly, I emphasize the foreign letters by including a foreign tag. For the next word, I did something similar to encode both the abbreviated form and the expanded form. For *sic*, it is the same, this time choosing between the original and the modern manner of writing.

This is as far as our discussion of TEI will go in this book. As with every tool and technology, for some situations, it will perform excellently and in another, it will not. What is especially troubling is that TEI can only be used for texts that are an 'ordered hierarchy of content objects,' meaning that its dissection needs no overlapping of tags. This restriction flows from TEI being a flavor of XML. It assumes that texts can (should?) be linearly ordered, and this excludes a lot of humanities research.[25] For example, in cultures with extensive commentary traditions, a text can be construed from an almost organic process of text reuse. To properly edit and analyze such a text, these other commentaries and base works need to be taken into account. For this, overlapping tagging would be necessary. Post-classical Islamic intellectual history is such a field. Islamic studies, in general, encounters another difficulty, namely the awkwardness of mixing an Arabic source text and its right-to-left script with English tags with its left-to-right script. Even if we disregard these problems, it still depends on the savviness and experience of the user to unlock the benefits of TEI. XML was originally conceived to be a form of markup that leverages the best of both worlds: discursive enough for human beings to read it, regular enough for machines to interpret. But the extensive apparatus that TEI prescribes makes it difficult for people to simply read and understand what is going on. We would need an interpreter, a piece of software that converts all the tags in visual markup, such as making titles bigger and bold and showing personal names in italics. This not only counts for reading but also writing. Writing TEI tags from memory is too difficult. However, solutions that provide automatic tag suggestion as soon as you type the first letters are labor-intensive,[26] and those that provide a set of buttons you need to click with a mouse to add a tag are ergonomically frustrating. This begs the question: if most of what we mark up by hand might be performed automatically in just a few years—when algorithms to detect names and dates are robust enough, for example—if so, is it not an absolute waste of resources to type all of it by hand? "Almost anything we do now, with painstaking effort, may be automated in five years' time,"[27] says musicologist Julia Craig-McFeely. This snarky remark should be kept in mind at every step of the way while marking up a text, especially when the choice has been made for TEI.

[25] The TEI Guidelines itself offers some solutions but these either cause bloat, inconsistency, a lot of extra labor, or all of the above.

[26] Such as in <*oXygen/*>, a favorite XML editor among Digital Humanists.

[27] Craig-McFeely, J. "Finding What You Need, and Knowing What You Can Find: Digital Tools for Palaeographers in Musicology and Beyond." pp. 307–39 in *Kodikologie Und Paläographie Im Digitalen Zeitalter 2*, edited by F. Fischer, Chr. Fritze, and G. Vogeler. Norderstedt: BoD, 2010, p. 330.

4 Intermezzo: Using the Right Editing Tool

If Microsoft Word (Windows) or Pages (macOS) or LibreOffice (Linux) are not the right environments to do our philological work, what is? The previous section already hinted at the correct but frustrating answer: it depends. And it depends not only on the research questions and the desired outcome but also at the moment in time since software development is constantly in motion. Several distinctly different use cases are discernible, I think, that will hold up regardless of the future technological development.

4.1 *You Need to Create a Small Critical Edition Solely for Print Purposes*

If there are no unusual features to the manuscript evidence you are working with, and all you need is to create a print publication, you might have enough with software such as Mellel or Classical Text Editor.

Mellel (macOS only) is a good example of an extended word processor.[28] It behaves similar to Microsoft Word but has several improvements that will facilitate scholarly writing. Above all, it handles large documents with significantly less trouble. Entire books or dissertations can be written on it as one document with virtually no lag. The find and replace functionality in Mellel is very advanced; it supports regular expressions and the option to save search queries. Cross-references throughout the document pose no problem and do not transform into a burden as the document grows. The same goes for footnotes, of which Mellel offers the ability to offer many different note streams, i.e., different sets of notes. For example, you can create a different series of notes for critical apparatus and a different one for comments. These streams can be in different scripts: Mellel has sophisticated support for blending scripts and writing directions. It is often the little things that push one piece of software far ahead of others. For example, with Mellel you can not only underline text (in more than a few ways, in fact), but you can also overline text. This is actually what is done in Arabic, so to be able to replicate that in an edition greatly improves the aesthetic experience.

The Classical Text Editor (Windows only) comes into play when a little more power is needed. It should be noted from the outset that its price-tag matches its learning curve: high. Since adoption is low, it is hard to get good training

[28] For my everyday writing I use Nisus Writer Pro, which supports Arabic, handles large documents without a problem, and has several useful extras such as a mode in which text is white on a black background, everything but the text is invisible, and the line on which you are typing remains in the middle of the screen, which is excellent to work without distractions.

material. Notwithstanding, it is, to date, the most complete and most advanced attempt at providing an ergonomic critical editing environment. CTE allows you to encode and markup the edited text, its variants, notes, translations, and anything else, within a graphical user interface without having to code or laboriously have to piece together the different bits of information. It does so by inviting you to store the different bits of information in different documents, while still seeing all those documents on your screen at the same time. This allows for elaborate critical apparatuses and extensive notes as you no longer feel bound by the dimensions of a piece of paper, as you do in Mellel. Technology, here, works for you in expanding your potential. This is equally true for CTE's export functions. You can transform this bundle of documents into one beautifully typeset PDF, one that rivals the looks of LaTeX, and you can also export it as a TEI-compliant XML file.

4.2 *You Are Working with a Very Large Text from Which You Will Mine Specific Bits of Information*

In this case, you should pay special attention as to how you will share your final data and what you want to publish out of it. In a nutshell, you could draw out your workflow, starting with the end, and take a few steps back—from the publishing phase to the analysis phase to the transcription phase—to see what technology you will use and how you can make sure that you have the easiest time using it. 'Mining' could be as easy as performing a search through the document for specific tags and then visually evaluating if the passage is relevant for your research or not. It could be as advanced as performing automated analysis using a programming language such as Python, where you will only look at the final statistics that it gives back. In either case, an approach like mARkdown is good. For this you can use a code editor like Visual Studio Code, Sublime, or OpenITI's favorite, EditPad Pro. Since the document will be too long and it will be difficult to go over it twice, you will have to be very sure what to tag so as to catch everything you need on the first run itself.

4.3 *You Have a Large Set of Separate Writings*

If you have something akin to a diary, collected letters or poems, or a journal or log, you will find much use in a VRE like Scripto. This will allow you to keep an overview of what has been transcribed and what has not, and you will be able (but not obliged) to recruit others to help you. You might simply transcribe the text without markup, or you could adopt a lightweight markup like Markdown, by which I mean the regular industry standard, not the specialized Islamic studies one. The advantage of a lightweight but broadly supported markup language like Markdown is that you will have a very easy time converting it to a webpage or printable document, as there are many applications that will recognize it and convert it to HTML.

4.4 *You Will Work Intensively with a Stable Text*

If your analysis is not so much concerned with variants across manuscripts copies but more about the structure and content of the text, and if the publication of the text itself is not interesting, it will be beneficial to use a stand-off annotation as implemented by TextFabric. In this case, you will work mostly in a code editor of your choice, using regular expressions to mold the text in the shape you want it to be. For example, you may obtain or produce a full transcript in plain text without any markup, replacing all the spaces with 'new line' characters to place every word (or, word group) on a separate line. If you are still working on your programming skills, you will rely on software like TextFabric, so it is crucial to abide by the required format of the software you use. Texts falling in this category are typically already published, but a heavily annotated or thoroughly reworked version might still be publishable, perhaps not as a print publication but as a digital data set. In this workflow, the information is maximally divided into different puzzle pieces, so it will take more effort to combine all these pieces together to get a printable document. The upside is that this workflow supports the heaviest and most detailed annotation; and therefore, there are no puzzle pieces missing to convert these different files into, for example, one TEI-XML file. The most obvious tool to put these puzzle pieces together is to write a custom Python script.

4.5 *You Wish to Critically Edit a Text with Unusual Features or Multiple Manuscript Witnesses*

If you have multiple manuscripts of an unedited text, or manuscripts with unusual features, and your primary intent is to publish the text as a critical edition (digital or print), word processors like Mellel or CTE will significantly hold you back. You will, in the long run, gain more from writing up your editorial work in XML format, for example, by using a subset of the detailed annotation system of TEI. You can do this in a code editor, or use a professional XML Editor. oXygen is a popular choice in the humanities, as it has support for TEI, meaning that when you start typing a tag, it will auto suggest whatever is available within TEI that starts with those letters. This makes typing in TEI faster and more precise.

You stand to gain the most from working with a subset of TEI that has proven its worth within your specific field. For example, *EpiDoc* has been highly successful in transcribing epigraphic evidence, and the *Scholastic Commentaries* *and Texts Archive* has worked on implementing TEI for medieval commentary traditions. The latter is an especially interesting example, as the person behind it, Jeffrey Witt, has created an entire toolset that includes functions to archive and publish your editions (both online and as a well-formatted printable PDF). This can be used freely. If at all possible, you want to forego reinventing the wheel, and for that reason it is vital to find out what technology has already been developed in your own field, or a neighboring one. In the case of *SCTA*, there are other benefits too, to adopt or at least be compatible with their standard. For example, this makes your edition interoperable with other editions. This means that if the text you work on cites another text within the scholastic corpus, you can include a direct link to that text (and a link back will be created, too). Whereas editing remains, in my view, the domain of the lone scholar, here we encounter great benefits from collaborating on the technical parts of our work. Indeed, basing your work on the toolset surrounding SCTA means that you need only very little skill with programming languages. It minimizes your time writing code or agonizing over how you will move from XML files to a printed edition, and puts the focus on actually producing the annotated transcription. Meanwhile, you are not locked in, so if you want to take your TEI-XML in a different direction, you can still do so. For example, if the manuscript evidence shows a great variety or only one or a group of witnesses attest a markedly different text, you could use web development technology to build a custom interface to include on/off buttons for different witnesses and arrange them horizontally or vertically, in different colors, to convey this information to readers.

4.6 *Concluding Thoughts*

What most choices have in common is that you are encouraged to work in text file formats using code editors or VREs. This separates data entry and analysis from publication. If you are unfamiliar with code editors, this may sound like a tall order, but advanced analysis demands advanced tools, and by separating raw data entry from polished publication, you have more freedom to mark up what you want. In addition, working in this way will ensure that you keep all your options open if you want to take your research in a different direction. For example, you may initially think you want to produce a critical edition, but later on, you may come to think of text mining analyses you would love to do. By having your raw transcription in a plain text format with regular markup, it will be very easy to reshape your text to prepare it for automated analysis. All of this will take varying levels of technical know-how, but by adopting the existing standards and tools, and with practice, you will minimize the time needed to learn or get trained in using those technologies.

5 Handling Images

Editing the text in a manuscript requires access to the manuscript and as part of a digital workflow it will be beneficial to have access to images of the manuscript. It is best to have those images in your private possession. Since there will always be more images not in your possession, you will need to outfit your workflow to also allow for externally hosted images. Likewise, if you can, it will be best to eventually share the images you used so that colleagues have a chance of verifying your analysis, similar as to how in the sciences raw data is made available to replicate the analysis. There is, therefore, a nearly constant flow of data between your own computer and others, mostly over the internet.

Pronounced 'triple eye eff,' the International Image Interoperability Framework is a rather elaborate and technical standard for image storing, sharing, and showing. Most of this technology is of no concern to us in the humanities: we can safely leave it to librarians, curators, and other specialists to make collections digitally available through IIIF. Nevertheless, knowing how IIIF works and having a passive knowledge of it, perhaps with practical experience in some regards, will be a great asset to your toolbox. This is not only true because you will encounter the IIIF standard a lot, but also because smart use of IIIF can provide some real benefits. For our purposes, IIIF introduces two technological feats: (1) images that are stored in different places can be loaded using a similar command, meaning that they can be brought together almost effortlessly, and (2) a basic set of manipulations can be directly requested when loading an image, such as only requesting a specific area or a specific quality of the image.

IIIF achieves interoperability and responsive loading by having users not directly querying the image file as it is on the server, but instead, users send a request to an API with some parameters, among them the ID of the image. In turn, the API does not use that ID to go directly to the image file, but it first looks at an abstraction of the image, called a manifest, which is a JSON file. Let us unpack those two sentences step by step, with the help of an example, to understand the conceptual framework that IIIF provides. If we return to our example of encoding "Christi autem …" and we want to have a photo of the manuscript folio of that sentence, we could simply fill in the direct URL of the photo on the website of the Walters Art Museum in a browser to get the image at the highest resolution possible.

Entering such a URL in a browser is what it means to request the image directly. This gives us all 91.7MB of the photo, and for storing the research data on your own computer, this method works just fine. But what if I want to associate the encoded text with the specific excerpt of that folio? What if I want just that fragment because it will be a smaller file? I could open the large file, cut out the line, and save it as a new image file, but this will produce a proliferation of new files, and unless I would give them a meaningful file name, it will be unclear what they actually are if they are copied and moved around. Besides, I might simply want to refer to the file as it is originally stored by the library or museum, as that will have a better chance of being persistently available online. IIIF's Image API is here to help us. API stands for Application Programming Interface and is, in general, a way for computers to talk to each other. You give a formulaic request to the API, and you get a reliable, formulaic response. For example, if you want to know a zip code, you could browse yourself to Google Maps, enter an address, and read it from your screen. You could also write a script (e.g., in Python) to look it up for you. Obviously, a computer cannot see, so it has no use of browsing maps.google.com. Instead, what it can do is send a request to maps.googleapis.com, specifying the address and asking to get a zip code back. Similarly, for referring to the manuscript image, we want to make a consistent, reliable reference, not just to the image as a whole but to an aspect of the image.

The Walters Art Museum does not support IIIF (yet), so our example needs to take a detour. This will only benefit us, as we can learn to take matters in our own hands and make use of IIIF more actively. Using IIIF to refer to an image means, unfortunately, that we cannot simply have an image stored wherever; it needs to sit on a server that runs the Image API. For simple purposes, this is easily done by uploading the image to Archive.org.

Using the new URL will make the image appear only slowly, as again the entire 91.7MB is downloaded. Archive.org, however, offers IIIF support if we go to a different URL and use the ID of the object, which is, in this case, 'W7000037600'.[29]

We now see the image through the IIIF Image API, and the first noticeable difference is that it appears almost instantly. This is because it is first offering us a dynamically created version of the image, which is very small, and once it is loaded, it will load a higher resolution in the background. In the top-right, we see a button 'Enable Cropper.' If we use it and select the region of interest, a pop-up appears with the desired URL. This functions as a query with several parameters, as such:

[29] A remnant of its origin: Walters MS W.7, f.17r, which must equal to the 37th photo, taken at 600dpi. Notice too that I provided some metadata on archive.org to describe the photo and link back to its origin.

/iiif/ID/REGION/SIZE/ROTATION/QUALITY.FORMAT

With the ID, we specify which among the many images on that server we are interested in. With the region, we can ask to offer only a cut-out. We do this by specifying the x,y-coordinates of the top-left corner and the bottom-right corner of a rectangle. Changing the size means the image you will get will have a different pixel-dimension. Rotation does just that: it will give you back (a region of) an image at an angle, which is very useful for marginal notes that are slanted or upside down. Quality can be changed to get a gray-scale or bitonal image. File format, finally, can give you an image in a format of your choice as long as the API can do the conversion. In this way, we ask the IIIF Image API to work on W7000037600 and only return us the region with the top-left point being (514, 3353) and the bottom-right point being (1270, 166), as big as the image allows, no rotation, in color, and as a JPG file (as opposed to the original, which is stored as TIFF). Notice, for example, how the resulting JPG weighs just 40KB, which is of course much easier to work with than the entire 91.7MB file.

Before the image file itself is requested, two other files are used: a manifest.json and an info.json. JSON is short for JavaScript Object Notation and is quite similar to XML, except it does not work with tags wrapped in angle brackets, but with "key: value" pairs. We will work with it in the next chapter. When you encounter an IIIF-compliant image, you can lift the curtains and look at what is going on behind the scenes, by requesting these files directly to look at their textual content. Archive.org automatically creates such a manifest file.

Let us write our own so that we get to understand the conceptual framework behind IIIF and experience using manifest files as a building block in our workflow. The one constructed for our example can be seen on the next page.

```
{
  “@context”: “http://iiif.io/api/image/2/context.json”,
  “@id”: “http://iiif.archivelab.org/iiif/W7000037600/
manifest.json”,
  “@type”: “sc:Manifest”,
  “attribution”: “The Internet Archive”,
  “description”: “Resources for Ch. 5 of <i>Among Digitized
Manuscripts.</i><div>Standards for Digital Editing or Editing
Digitally.<br /></div><div>Page and excerpt taken under CC0
license from Walters Art Museum, available at\u00a0https://
manuscripts.thewalters.org/viewer.php?id=W.7#page/1/mode/2up</
div>”,
  “label”: “Among Digitized Manuscripts Chapter 5 Standards
for Digital Editing or Editing Digitally”,
  “logo”: “https://encrypted-tbn2.gstatic.com/images?
q=tbn:ANd9GcReMN4l9cgu_qb1OwflFeyfHcjp8aUfVNSJ9ynk2IfuHwW1I4mD
Sw”,
  “metadata”: [
    {
      “label”: “title”,
      “value”: “Among Digitized Manuscripts Chapter 5
    Standards for Digital Editing or Editing Digitally”
    },
    {
      “label”: “subject”,
      “value”: “Digital Humanities”
    }
  ],
  “seeAlso”: “http://archive.org/metadata/W7000037600”,
  “sequences”: [
    {
      “@context”: “http://iiif.io/api/image/2/context.json”,
      “@id”: “http://iiif.archivelab.org/iiif/W7000037600/
canvas/default”,
      “@type”: “sc:Sequence”,
      “canvases”: [
        {
          “@id”: “http://iiif.archivelab.org/iiif/W7000037600/
canvas”,
          “@type”: “sc:Canvas”,
          “height”: 5948,
          “images”: [
            {
              “@context”: “http://iiif.io/api/image/2/
context.json”,
              “@id”: “http://iiif.archivelab.org/iiif/
W7000037600/annotation”,
              “@type”: “oa:Annotation”,
              “motivation”: “sc:painting”,
              “on”: “http://iiif.archivelab.org/iiif/
W7000037600/annotation”,
              “resource”: {
                “@id”: “http://iiif.archivelab.org/iiif/
W7000037600/full/full/0/default.jpg”,
                “@type”: “dctypes:Image”,
                “format”: “image/jpeg”,
                “height”: 5948,
                “service”: {
                  “@context”: “http://iiif.io/api/image/2/
context.json”,
                  “@id”: “http://iiif.archivelab.org/iiif/
W7000037600”,
                  “profile”: “https://iiif.io/api/image/2/
profiles/level2.json”
                },
                “width”: 5137
              }
            }
          ],
          “label”: “Among Digitized Manuscripts Chapter 5
Standards for Digital Editing or Editing Digitally”,
          “width”: 5137
        }
      ],
      “label”: “default”
    }
  ],
  “viewingHint”: “paged”
}
```

Here, we see that the entire document is wrapped in curly brackets. This is a JSON-way of indicating that all the information belongs together. In the case of an IIIF manifest file, the information belongs together because it describes one real-world artifact. The first two lines are necessarily included to alert any person or computer reading this file that it should be read as an IIIF manifest. The next lines, up to "sequences," describe the artifact in its general characteristics. ID is necessary so that, if we were to open multiple files at once, we could always distinguish this one from other manifest files. "Label" is a more human-friendly version of the ID. Inside "metadata" we see square brackets. This indicates that the value we want to associate with the key "metadata" does not consist of just one piece of knowledge, but multiple ones, each of them wrapped in curly brackets and separated by a comma. In this case, we only included two items, but ideally, a full codicological description is given here. We leave it out here so that the entire manifest file can fit on two pages and also because this description is already offered in a machine-readable format by The Walters Art Museum, namely in a TEI-compliant file called "W7_tei.xml." A reference to it is given under "seeAlso," together with some hints to read this file as a TEI-XML.

What we discussed so far is for IIIF what "teiHeader" is for a TEI document. Just as "text" comes next in a TEI file, what comes next in an IIIF manifest is a definition of the actual image content. To do this flexibly, images are not simply defined but painted onto a canvas. A canvas is an abstract rectangle, defined in IIIF by making its type "sc:Canvas," given it an ID, and defining its width and height. One manifest can include many canvases. For example, an entire book can (and usually should) be defined by one manifest in which each page is defined as a canvas. The image of the page is only defined within the canvas, painted onto it, with the ID in its "resource" equal to a URL that leads to an IIIF Image API, including its usual parameters. In most cases, we request from this Image API the full image in its best quality, without rotation, and paint it onto the canvas such that it fills up the entire canvas. It is, however, entirely possible to have the image only fill up a quarter and you can define three other images to fill up the other quarters. Since the images can come from different locations, IIIF gives you the potential to reconstruct a page if the torn pieces have been digitized by different institutions, or to reconstruct a manuscript if different folia are scattered across multiple libraries.

You will have noticed that the images are wrapped inside canvases, which in turn are wrapped in sequences. A sequence defines the order of canvases. For example, for a manuscript for which it is known that pages have been reshuffled, you could define a sequence of canvases in the order of which the artifact currently is, and a separate sequence in which you define canvases in the order in which you think the manuscript ought to have been in the past. You can even include empty canvases to represent missing pages.

With its nesting, a manifest JSON can be a bit overbearing to look at at first, but with the conceptual framework in mind, it need not be hard to read it. In our case, we have defined one sequence, with one canvas of 5137 pixels wide and 5948 pixels high, and on that canvas we have defined one image.

IIIF also allows references to text annotations for each canvas. Thus, we could have included a reference to our TEI encoding of “Christi autem....” I have thought it sufficient to include a reference to this IIIF manifest within our TEI file. After all, the TEI file is dependent on the IIIF manifest and not the other way around. At the moment of writing, there is no built-in support in TEI or IIIF for the other standard such that it truly leverages the strengths of both into a powerful combination. It seems reasonable to expect that the organizations behind each standard will work on this in the coming years. Since the combination of digital image surrogate and digital text surrogate is of paramount importance for anybody working with manuscripts, we will have to follow these developments closely.

Having reading knowledge of manifest files is a great asset. For example, if you encounter an IIIF viewer (such as Universal Viewer or Mirador) that shows a digitized book that allows for text searching, you know that the manifest file will contain links to images of all the pages as well as one or more links to plain text files containing (for example, an OCR version of) the text. You can open the manifest.json in a code editor like Visual Studio Code to extract those links. Upon opening, the file might just be sitting on one single line, but searching for “http” will quickly find you links to images. Studying those links, you can find out what extension the images preferably have, for example .jpg. If so, you can search, with the option for ‘regular expressions’ turned on, for “http.*?jpg”. Regular expressions allow you to search more flexibly than just using a string of characters. For example, a period represents any character, and a star means as many of the foregoing. Therefore, “.*” means ‘as many characters as possible,’ and adding the ‘?jpg.’ means that as few characters as possible are found between the ‘http’ and the ‘jpg.’ This, then, makes you find entire URLS for all images. By clicking Select > Select All Occurrences, you will have selected all these URLS, which you can now cut and paste in a new file. Since we know the parameter settings of the Image API, we can make sure the region and size are full, and the quality is native for all images to get the best possible image. By creating a clean list of URLS, we can then download them, either with a script, command line tool,[30] or an application with a GUI.[31] For digitized manuscripts that are not served through an IIIF Image API, your best strategy is to find a regularity in the URLs for the images of successive pages and write a script to download files for which the URL is adjusted every time according to those regularities.[32] If you cannot right-click and 'Open image in new tab' to see the URL, you can probably still find it by digging into the page resources.

6 Archiving and Publishing

We previously defined the workflow for digital editing as follows: Digitizing → Transcribing ⇄ Analyzing → Publishing. This book focusses on the first half of that workflow and the parts of the Analyzing-phase that are included in this book focus on what depends on manuscript images and not the text per se. The reason for this is that there are already plenty of resources regarding digital approaches to text collation, stemmatics, reception analysis, semantic analysis, topic modeling, how to build a who's who or a gazetteer, or other fascinating pieces of analysis that go along with philology. Most of these discussions combine analysis and publication into one, under the term 'digital edition.' This term is a tender topic in DH. Everybody is highly opinionated and passionate about what we ought to do. Before we finish this chapter with a discussion of digital editions, let us first delve into the essential step of archiving.

Archiving digital materials means storing a set of files together in a permanent state in the double sense of the word: permanent as a guarantee that the files will not change over time, and permanent as a guarantee that the files will remain available over time. The latter sense can be further split out into two components: available in the digital world can mean 'to persist on a hard drive,' and it can mean 'to be accessible to a wider audience.' This makes archiving easier said than done. Regarding the first sense of permanent, you face a conundrum since you want to continue to work on it. Archiving would mean to preserve its state while working on it would mean altering it. As for the second meaning of permanent, you certainly want to share it publicly so that others can verify your analysis, but you only want to do so to the extent that is legally

[30] A quick "wget -i listOfURLs.txt" would do it.

[31] For example, on macOS, you can paste those URLs in TextEdit and right-click and select Services > Download links with Folx, which gives you a graphical user interface to download multiple files.

[32] With Python, you can use the package urllib to download items on the internet.

and ethically allowed. Further, if you intend to keep working on it, which is likely, it is only normal to make sure you are not oversharing so that nobody will beat you to the publication of exceptional insights that the data provides. There are currently no clear solutions to these problems.

For your private use, you should consider adopting the 3-2-1 rule: all your truly important data should be stored in three different places, on at least two different kinds of media, of which at least one is off-site and preferably offline. Next to that, you can think of your data in terms of circles of influence, with a core set of files well protected and in immediate reach imagined as a circle, and rings of ever less important files less protected and more laborious to access. Since you likely do not want to spend a lot of time and money on this, your scenario will probably be a variation of the following: a core set of files that make up your current projects sits on your computer in a folder synced with a consumer cloud storage solution.[33] Since you occasionally backup your computer, you are already compliant with the 3-2-1 rule. In fact, if your computer has an SSD hard drive, and the external hard drive for your backups is a regular one (with a spinning platter), you have no less than three different media, as cloud storage can be considered a different kind in this case. It is likely that your next ring consists of primary sources, secondary literature, and reference materials that you value highly or otherwise use frequently. They can sit on your computer, but are not synced to the cloud. Instead, you can choose to mirror them on an external hard drive (other than the one you use for backups). On this external hard drive, you may store additional sources, books, articles, and manuscripts that you would not use often. This external hard drive, in turn, could be occasionally synced with another external hard drive or a network attached storage unit on an offsite location. This way, your most valued literature also complies to the 3-2-1 rule. The other files do not, since they are not stored thrice. If you move around a lot, you may want to consider to change the cloud option for an SD card or USB stick, which you can carry with you.

With this need for different storage devices and the likelihood that you will sometimes exchange such devices with colleagues, you need to know of the standard controlling the normal operation of a storage device. This is a 'file system.' For Apple devices, the file system is called APFS. If you use Microsoft Windows, however, you will likely have your drives configured (even without you knowing or choosing) for NTFS. Linux users, meanwhile, probably have ext4 as their file system. These file systems do not mix and match. NTFS, for example, cannot be used on a Mac. That is, you will be able to read files, but not able to write files. Third-party (paid) software will help you here, but you could

[33] Such as Microsoft OneDrive, Google Drive, Apple iCloud, Dropbox, and SugarSync.

also format the disk to another file system. If you intend to share data with colleagues or friends regularly, you would not want to have such restrictions. The solution is to format the device to the file system called exFAT. Most versions of Windows and macOS can handle this, and it can even count on some support on Linux. exFAT is a file system which can take care of files of practically any size with practically any file name, organized in practically any tree of folders.[34]

Public archiving is another task. Your first port of call should be your own university, but this can currently be a drawn out process, where you might not get the type of support you are looking for and, frankly, not a long term assurance of hosting. Next, you can look for what else is going on at a national level or in your field of studies. Perhaps, a digital text corpus or other kind of repository is forming, to which you can contribute your data of your critical edition. The upside of this is that this combats 'silofication,' which is the danger of various digital resources living in ignorance of each other, whereas their connection could provide a much more powerful resource. As a corollary, it could increase your exposure since more people will normally go to the central repository and will find you through it. Moreover, when this repository finds traction, it will become an acceptable point of reference to which you can refer in notes in publications. The downside is that you are obliged to follow the rules and standards of whoever is in charge. If they do not update the corpus often, or want to format the texts in a specific way that may be useful for their own needs but not for others, its normal usability is jeopardized. As long as you do not (or frankly cannot) make meaningful use of the corpus already, it is unlikely you will contribute to the repository.

What can you do on your own? An unusually good option has been born: Zenodo. Through this service, you can create a free academic repository that the developers claim is guaranteed to be saved for many years to come. Zenodo allows virtually all the things one would want in a repository: you can store the data, the metadata describing it, and create a DOI, Digital Object Identifier. In Chapters One and Two, we noticed that one pesky problem of the digital world is that it does not have a stable reference system like the print world, with its publication information and the page number. This DOI is the digital world's answer to that problem, meant to be a persistent and precise referent to a book, article, edition, or any piece of digital data. A similar initiative to refer to one and only one person is ORCiD. An account on the latter is free, and

[34] To format a device using a Windows computer, connect the device and use the program *Disk Management*. On macOS it is called *Disk Utility*. For Linux, there is not a standard program with a point-and-click graphical user interface (GUI), but *GParted* is trusted and used a lot.

Zenodo is a way to get a free DOI for your data. Most publishers assign a DOI for your publication. Therefore, you can link the DOIs of your publications and repository data to your ORCiD and vice versa to ensure that people understand you created these documents and that these documents were created by you. A good repository should include as much of your data as possible, including images, transcriptions, markup, and code. Ideally, there should be some documentation within the file or in the filename to make clear what the file is and what it does. A readme file should come along with the bundle of files, explaining the characteristics and purpose of the data set as a whole. This readme file should also include some information about creation and ownership. Either within the readme or as a separate file, you should include a license that describes what other people are allowed to do with it. Most commonly you will simply take over a license from, for example, Creative Commons. Lastly, it is best practice to include more documentation describing how the data can best be used.

If you wish to be sure and store things twice, or you do not wish to use Zenodo, there are alternatives. For plain text files such as programming code, marked-up transcriptions, notes, and statistics, there are dedicated services whose benefits consist of functionality, such as version control, easy duplication and reworking, and, after that, reintegration. The most prominent example is the free service GitHub, discussed in the next chapter. Other data, such as images and other binary files, can be uploaded to The Internet Archive. The functionality it offers is very limited and is geared at dumping the final archive onto it. What you get in return is storage at a permanent URL, with one of the strongest guarantees of permanent worldwide accessibility.

Should you publish a digital edition of your digital files? You will have to answer that question yourself, but I can summarize the discussion to help you make that decision. For a broad overview, you can best look elsewhere.[35] Practical introductions are mostly in German.[36] A good practical introduction in English

[35] I have enjoyed the summarizing discussion in a dissertation: Hörnschemeyer, J. "Textgenetische Prozesse in Digitalen Editionen." PhD dissertation, Universität zu Köln, 2013.

[36] Sahle, P. *Digitale Editionsformen: Zum Umgang Mit Der Überlieferung Unter Der Bedingungen Des Medienwandels*. 3 vols. Norderstedt: BoD, 2013; Jannidis, F., H. Kohle, and M. Rehbein. *Digital Humanities: Eine Einführung*. Stuttgart: J.B. Metzler, 2017; Kurz, S. *Digital Humanities: Grundlagen Und Technologien Für Die Praxis*. Wiesbaden: Springer Vieweg, 2015;

is Apollon, Bélisle and Régnier's *Digital Critical Editions*.[37] An excellent theoretical discussion, one that continues where this book stops, is Pierazzo's *Digital Scholarly Editing*.[38] The discussion surrounding digital editions can be succinctly summed up in the following four questions. (1) Now that our workflow happens in the digital world, and we have witnessed innovative ways of doing our research, can (or, should) we find new ways of publishing our results? (2) If we do so, when is our product 'digital' enough to be called a digital edition? (3) If we have made it digital enough, at what point is our product truly an 'edition'? (4) If we make truly digital editions, how can we make them count just as 'critical' as print publications?

The first question oscillates between *can* and *should*, and this is because those in DH who work on this topic are so enthusiastic of the possible innovative, technology-driven scholarly products that they go beyond the mere possibility and argue that the digital era has dawned and we would be foolish to hold ourselves back by the paradigm of the print world. Thus, Burdick et al., authors of the manifesto-like *Digital_humanities*, boldly write: "We are moving […] to an era of scholarship based on the collaborative authoring possibilities of the 'great project',"[39] and a little later claim that matters of publishing "have moved to the forefront of the digital humanities precisely because choices of interface, interactivity, database design, markup, navigation, access, dissemination, and archiving are all part of how arguments are staged in the digital world."[40] They claim, then, that a change in scholarly output is inevitable since these aspects are ingrained in the fabric of the digital world. Similarly, Van Zundert writes: "it cannot be the primary intent of a digital scholarly edition to represent a scholarly edition that is better represented by a print publication." And he continues to argue that a digital edition should exploit "the idiosyncrasies of the digital environment."[41] Robinson, meanwhile, calls digital editing a revolution and sees the revolution, especially play out in "changing what we

[37] Apollon, D., C. Bélisle, and P. Régnier, eds. *Digital Critical Editions*. Urbana: University of Illinois Press, 2014.

[38] Pierazzo, E. *Digital Scholarly Editing: Theories, Models and Methods*. Farnham: Ashgate, 2015. My references are to the online, open access version.

[39] Burdick, A., J. Drucker, P. Lunenfeld, T. Presner, and J. Schnapp. *Digital_humanities*. Cambridge Mass.: The MIT Press, 2012, p. 83.

[40] Burdick et al, p. 84.

[41] Zundert, J.J. van. "By Way of Conclusion: Truly Scholarly, Digital, and Innovative Editions?" pp. 335–346 in *Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches*, edited by T. Andrews and C. Macé. Turnhout: Brepols, 2014, p. 338.

make."[42] He reflects on what a critical edition is and says: "all this proclaims 'I am a hypertext: invent a dynamic device to show me.' The computer is exactly this dynamic device."[43]

The previous question already hinted at the high standards these authors want to set for such editions. Robinson spurns features such as browsing and searching through a text: "you can do these things with print editions too."[44] For a digital edition to be digital, Robinson seems to want a) all manuscript witnesses transcribed individually, b) all transcriptions to be in both original shape and regularized, c) all manuscripts available as digital photos, and d) the ability to change or add to the edition by multiple people, preferably anyone. Van Zundert argues for a nearly identical list: a digital edition ought to have a) any extant data including all transcriptions and facsimile images and any other notes, b) additional media such as sound and video, c) full-text search, and d) the ability to change or add to the edition in a collaborative fashion, possibly open to anyone. From the previous quote of the authors of *Digital_humanities,* we can infer they would agree with such a list.

Do we still end up with a critical edition if we would follow up these demands? Let us structure the answer to this question as counterarguments. First, to wish for full transcriptions and markup seems to wish for an objective record of the facts. Next to the question whether all these facts really need recording, we may note that transcription itself is not an objective task but can be considered an editorial task dependent on the editor's erudition.[45] Second, Van Zundert's invitation to make the edition a richer experience with mixed media begs the question: why? Certainly, in scholarly editing of premodern texts, illustrations, sound, and video are hardly ever needed. It makes no sense to include them for the sake of inclusion. Third, there is a strange sense of demanding ever-more realistic representations of the original documents, ending in demand for facsimile photos. The point of an edition is exactly to move beyond the manuscripts in order to facilitate easier access. The advocates of digital editions laud the democratizing nature of the digital world, claiming that now manuscripts are not exclusively for the scholarly elite.[46] But restricting access, I would contend, was never the point. Whereas certain documents

[42] Robinson, P. "The Digital Revolution in Scholarly Editing." pp. 181–207 in *Ars Edendi Lecture Series, Vol. IV,* edited by B. Crostini, G. Iversen, and B.M. Jensen. Stockholm: Stockholm University Press, 2016, p. 198.

[43] Robinson, P. "Current Issues in Making Digital Editions of Medieval Texts -Or, Do Electronic Scholarly Editions Have a Future?" *Digital Medievalist* 1 (2005).

[44] Robinson, "The Digital Revolution ...", p. 193.

[45] Pierazzo, p. 83.

[46] Burdick et al, pp. 80, p. 87; Robinson, p. 198.

from the early modern period can be understood and found interesting by the general public, this is simply not the case for the vast majority of manuscripts, which require special training to be deciphered and a long study trajectory to be found meaningfully interesting. The task of editing such texts is exactly to make that training unnecessary and that trajectory considerably shorter. An example Robinson likes is Prue Shaw's edition of Dante's *Commedia*, which he lauds for not actually providing an edited text. "This absence strikes me as the single most remarkable element of the edition," he says.[47] While there certainly is merit in what he argues for, the unwillingness to let the editor make a scholarly argument in favor of a particular reading means we head in the direction of a mere repository. Apollon et al., in their *Digital Critical Editions*, argue specifically against Robinson. They make the obvious but highly pertinent remark that throwing together a bunch of transcriptions with markup and images "does not facilitate reading,"[48] and explain that "quantity rapidly becomes a problem to manage at both the level of display as well as the level of codification." This, of course, goes against the very nature of editing. Editing is about deciding which particular bits to pick out of the rich tapestry of the actual evidence to make the original text maximally understandable with minimum distractions. Simply passing on all of it is not helpful. Pierazzo makes a good and precise point when she says that "the easier it becomes to publish on the Web, the larger the quantity of textual materials available, the more important becomes the guidance of the editor. Quality still matters."[49]

Quality does indeed matter, and it matters the most if you want your edition to be taken seriously. What Robinson and Van Zundert argue for, however, puts quality in doubt in two ways. For one, by allowing anybody to contribute to the transcription (crowdsourcing), quality is hard to ascertain and even harder to be equal across the document. Further, by allowing changes after the first release, quality can change over time. Even though these aspects make a digital critical edition typically digital and do not jeopardize its nature of being an edition, they are exactly those aspects that make them less critical. Pierazzo points out that if an edition can change over time, reviewing becomes meaningless.[50] It seems that scholars know this intuitively, as there seems to

[47] Robinson, "The Digital Revolution …", p. 196.

[48] Apollon et al, p. 21.

[49] Pierazzo, p. 8. Note however that Pierazzo nuances (perhaps contradicts) herself elsewhere when she makes the point that "digital editions based on text encoding allow the luxury of not choosing." cf. Pierazzo, E. "Textual Scholarship and Text Encoding." pp. 307–321 in *A New Companion to Digital Humanities*, edited by S. Schreibman, R. Siemens, and J. Unsworth. Oxford: Wiley-Blackwell, 2016, p. 313.

[50] Pierazzo, p. 143.

be hesitation in citing digital editions. Indicative is Melissa Terras' answer to her own question regarding how a digital edition project, *Transcribe Bentham*, can benefit her: not by contributing to it but by using it in conventual paper publications.[51] Robinson himself admits that electronic publications are not seen as 'real.'[52] Burdick et al. think that DH should come up with new "evaluative metrics for legitimizing and credentialing this kind of scholarship,"[53] but this will be exceedingly hard given that digital editions are often very different from each other. This additional aspect makes digital editions odd indeed. For how can one evaluate something if it is unique and incomparable? Print publications have much more in common with each other, and this makes for easier comparisons as the format itself is reliable. It might just become a factor at the most crucial moments of your career when your output is scrutinized and compared to others because you are up for a job or tenure.[54]

Overseeing the pros and cons of digital critical editions, it seems as though we have a 'triple constraint' going on, in which we have 'digital,' 'critical,' and 'edition' and can only pick two. It would be a good thing if we saw a tandem model emerge, in which a scholar or a team of scholars produce both a (paper) critical edition and a digital (less critical) edition, especially in cases where aspects such as multiple layers of text reuse and internal referencing play a large role, which can be made much clearer through a digital edition. Such a model could foster a conversation to figure out exactly how a digital critical edition can fit into our discourse. Experimentation is crucial here. With the things discussed in this chapter, together with the next chapter, you should be able to pull together a web-based edition and experiment for yourself.

[51] Terras, M. "Present, Not Voting: Digital Humanities in the Panopticon." pp. 172–90 in *Understanding Digital Humanities*, edited by D.M. Berry. New York: Palgrave Macmillan, 2012, pp. 179–180.

[52] Robinson, "Current issues ..."

[53] Burdick et al, p. 84.

[54] Cf. Prescott, A. "Consumers, Creators or Commentators? Problems of Audience and Mission in the Digital Humanities." pp. 61–75 in *Arts & Humanities in Higher Education* 11, no. 1–2 (2011), p. 62.

CHAPTER 6

Cataloging: From a Dusty Backroom to the World Wide Web

Owing to the mere fact of their accessibility, it has been established that digitizing little used, obscure materials can increase their usage. In the case of a collection of American literature from the nineteenth century at Cornell, the increase of use was so dramatic that it was concluded that “In hard copy the material may have seemed obscure; when digitized it becomes a core resource.”[1] Manuscript collections only benefit more from this effect, as their items are unique and only to be handled under strict regulations. Then again, mass digitization will only make a real impact if it goes along with mass metadata, a digital catalog of some sort, with information like title, author, dating, and provenance.[2] In this sense, creating digital catalogs should be a priority. For the most part, this is the domain of librarians and curators, who are experts at this just as much as they know far better how to do the digitization than we, scholars, do. But just as we digitize on a small scale ourselves, when working on a research project, we will often find ourselves in a situation where it would be beneficial to create a small catalog for our private research needs. It should be noted that ‘catalog’ here can refer to anything as long as it is a collection of items described under certain terms. So, just as we can catalog books, we can also catalog illuminations, glyphs, or abbreviations. In this chapter, I present a case study to explore how we can do field work digitally and transfer the labor of that fieldwork into catalog data and then how, from that catalog data, get to a digital catalog accessible online, much like the online catalog developed by a team at the University of Trier of the collection of the German monastery of St. Matthias, but then developed by ourselves and without a budget. Much of

[1] Hirtle, P.B. “The Impact of Digitization on Special Collections in Libraries.” pp. 42–52 in *Libraries & Culture* 37, no. 1 (2002), p. 43.

[2] Nichols, S.G., “Materialities of the Manuscript: Codex and Court Culture in Fourteenth-Century Paris,” pp. 26–58 in *Digital Philology: A Journal of Medieval Cultures*, vol. 4, no. 1 (2015), p. 27; Ornato, E. “La Numérisation Du Patrimoine Livresque Médiéval : Avancée Décisive Ou Miroir Aux Alouettes ?” pp. 85–115 in *Kodikologie Und Paläographie Im Digitalen Zeitalter 2*, edited by F. Fischer, Chr. Fritze, and G. Vogeler. Norderstedt: BoD, 2010. p. 96; Riedel, D. “How Digitization Has Changed the Cataloging of Islamic Books.” *Research Blog Islamic Books*, August 14, 2012. Dahlström, M. “Critical Editing and Critical Digitisation.” pp. 79–98 in *Text Comparison and Digital Creativity*, edited by W. van Peursen, E.D. Thoutenhoofd, and A. van der Weel. Leiden: Brill, 2010, p. 90ff.

this work rests on web development technology, which is an especially potent part of computer technology for many parts of our work in the humanities.[3] Importantly, web development is very easy to learn since there are a plethora of resources freely available. At the same time, one significant downside to web development is that it is undergoing an extraordinary evolution, with new major innovations pushed out almost every year as new standard ways of working. Therefore, in this chapter, we shall focus on the fundamentals in the understanding that with this knowledge, you will be able to go out and adopt any new technology that might benefit you.

1 Field Research Workflow: From a Dusty Backroom to My Computer

In this section, I shall describe the workflow I settled on when I worked on cataloging a small collection of books, articles, and manuscripts. These artifacts are kept at Sankt Florian, a monastery near Linz, Austria. Sankt Florian, in its current form, was built in the 17th century in a baroque style. It is basically a monastery and palace in one. One part of it was for the Augustinian Canon Regulars to live and pray, while the other for the Habsburg Monarchy to stay for the night and conduct business. St. Florian is a center of music with a world-famous organ, but also boasts a very large collection of books, part of which are kept in a dramatic baroque gallery. On their website, it says that:

> In 1930, the Monastery library bought the literary remains of Rudolf Geyer (1861–1929), a Viennese orientalist. Still 20 years later, this collection was considered the most comprehensive one in Arabic literature between Berlin and Rome. Meanwhile, about a third of Geyer's books is indexed.

Naturally, I was intrigued. I wondered exactly how big a 'comprehensive' collection it would be, and I figured that given the life years of Rudolf Geyer, there must be rare or otherwise valuable books from the 19th century in this collection. After inquiring, the 'index' of one-third of the collection turned out to be a handwritten list of title, author, place, and year, only to be consulted in the library itself. Further details about the collection could not be given. An on-site visit was unavoidable, which I did in the summer of 2016. Arriving late on Monday and leaving early on Friday, I only had three full days to conduct

[3] Susanne Kurz rightfully made it one of the pillars in her practical introduction to Digital humanities, see Kurz, S. *Digital Humanities: Grundlagen Und Technologien Für Die Praxis*. Wiesbaden: Springer Vieweg, 2015.

a preliminary investigation. As a testament to the power of a digital work environment, although the catalog was only completed two years later, it was all based on my fieldwork of just those three days.

Upon arrival, I came to know that the Geyer collection was not in the beautiful gallery part of the library, but in a dusty back room. It covered perhaps about twelve bookcases, each with seven shelves, many of them containing double rows of books. Inspecting the hand list that had been drawn up, I noticed that the previous cataloger had looked at certain sections that contained only European language resources, of which most of them were articles, individually bound and shelved. Browsing for half an hour made it clear that the majority were books. There were also a lot of articles as offprints, and a dozen or two manuscripts. In short, anything from a book review to a multi-volume primary source could be an item, with each one having a seal, like a post stamp with an index number. Interestingly, inside virtually all items, Geyer placed an Ex Libris sticker, which also shows a number, different from the one on the seal; thus, apparently, there are two numbering systems. In cataloging the collection, I only considered the index number on the seal, since this is on the outside of the item and is, therefore, easier to inspect while browsing the shelves. Next to this collection were boxes with notes, drafts, letters, and other things belonging to Rudolf Geyer. I did not investigate them in any detail, focusing on the proper collection.

Sitting down for each item of the collection and noting the catalog details would have been too time consuming. Even if I had much more time at the monastery, it seemed like an ineffective workflow. Given the size of the collection (about 1500 items) and the time left, I decided I could make a photographic index of all title pages, or at least of those items not mentioned in the hand list. I used my phone for this, an iPhone 6, which facilitated my process. Even while holding the phone in my hand, I could use both hands to pick up items and flip them open. Keeping them open with one hand, I snapped a picture with the other. For lower shelves, I used a trolley to load part of a shelf onto it, returning the item after taking a picture. This was also necessary to reach the second back row. On the top shelves, which had a single row, I simply stood on a ladder, using its top surface as a small table to keep the items straight while photographing them.

I did not use the stock camera application but Evernote.[4] Evernote is a simple note-taking application. It is essentially a database for notes of all kinds, typed text, drawn handwriting, audio, images, and PDFs, with a user-friendly

[4] There are alternatives. For example, Tropy is free software specifically developed to take in thousands of photos a researcher takes at an archive and provide the user with a friendly way

interface built around it that is so polished and easy to use you rarely think of it as a database, but simply as a note-taking app. With an eye towards possibly reusing your current labors, I would recommend using such an application. It keeps all your notes together and stores them in a way that will be accessible and exportable for the long-term foreseeable future. Notes are generally inserted in different notebooks, but with one search, one can find notes across all notebooks. It also ensures an offsite backup, as everything is stored on Evernote's servers too. In the case of cataloging Geyer's collection, it allowed me to store photos of each item in separate notes, so that later on I would never have to doubt whether a photo belongs to one or the other book. Also, if I want to export the photos to make them ready for another application, a rudimentary division is already baked in. If I ever wanted to bring all the photos together, it would be possible too. Originally, I wanted to give each note the title of the item number as it is written on the post stamp-like seal. After only a few items, I realized that typing them out was taking up too much time. Instead, I opted to simply make another photo, this time of the seal. In certain cases, for example when there were multiple title pages in different languages, or when there were multiple volumes of one title, or when parts of the catalog information was not on the title page but on the last page, I snapped additional pictures. In total, I took probably around 2,500–3,000 photos. If we assume three full days of eight hours work, that would come down to half a minute per photo, or about a minute per item. According to my notes in Evernote and comparing their time of creation, I did indeed spend a minute—sometimes even less—per item.

However, snapping so many photos in Evernote came with a cost. To upload everything to Evernote's servers took over a week. Every photo taken in Evernote is also saved in your Photos app. Since all applications on an iPhone are sandboxed, the photos are physically stored twice on my phone. The upside of having the photos also in the Photos app is that I was able to quickly put them on my laptop. I only had to connect my phone, and the Photos app on my Macbook appeared, allowing me to select and download the images. The downside is that, after, having the photos twice on my phone seemed redundant, but trying to delete the photos from my Photos app was surprisingly difficult. Apparently, deleting two to three-thousand photos at once from Photos is an incomprehensible task for the phone. I tried it several times but failed. I also tried to do it for each day, basically in batches of about eight-hundred, but that too resulted in nothing. The iPhone simply remained unresponsive and

of making sense of all the photos back at home. A more fully-fledged alternative to Evernote is Onenote.

did not delete anything. I ended up deleting them in a hundred or so batches of twenty photos. It is entirely possible that such issues are resolved in the future, but undoubtedly, you will encounter other but similarly odd behavior. It seems, then, that we are stretching the capabilities of consumer electronics and stock apps. At the same time, it is pleasant to notice we can actually get by with these simple tools, and we do not need to acquire more professional hardware or software to do our job.

By then, I had the title pages of the Geyer collection right in my pocket, stored in Evernote. Considering the number of notes I ended up with, together with the written hand list, I estimated that the total number of items amounted to no more than 1,500. The next step was to extract the different elements of the title page (title, author, publisher, etc.) into plain text, collected in such a way as to be able to be constructed into a catalog. I considered if I could make the jump from images straight to professional, library-quality cataloging. This, however, was unnecessary. Libraries use database systems that require vast amounts of entries, constant updating, and write-access for multiple users. None of that applied to the Geyer collection. All I needed to end up with was a machine-readable list of all the items with their details so that we can then reuse the list in different ways; either to create a printed catalog or an interactive, online catalog or to load it into a bigger catalog. This list would contain only a limited number of entries, at least from a computer processing point of view. Furthermore, the list would require little to no update afterward so the ability to edit entries did not need to be fast and user-friendly. Lastly, I knew I was the only one who was going to put in hours of work into this, so there was no need to allow multiple users.

To reduce the hours of work needed and make the actual work as painless as possible, I considered making a custom application in FileMaker. This is a software with which you can create simple relational databases. Using its drag-and-drop elements, you can create forms to either display or enter records. A relational database is best visualized as several tables held together by relations. In each table, a row indicates a record, representing a unique object that is described by the values written in that row in several columns. For example, a table *persons* can have columns such as *first name, last name,* and *age*. Each row then represents a person, described by their name and age. This person is unique: it is only defined once in the table. It should be noted that a table is only a visualization. When we look at tables, there seems to be a specific order, with a top and a bottom for the rows and a left and a right for the columns. For databases, this kind of order is not actually there: all records are stored as though they are marbles in a bag in which you put your hand to reach for them blindly.

In this example, if you also wish to include a column *books*, to describe all the books the person wrote, you will encounter a problem. For some people, there will be no books; for some, one title; for others, multiple. We would ideally have a dynamic number of columns to fill the number of books per person. Let us assume for argument's sake that each book has only one author. Then, a better way to write this down is to open a new table called *books*, listing each unique book as a row with columns for *title* and *author*. In the author field, we only need to put a referral to the correct row in the *persons* table. Such a referral is called a key, a foreign key to be exact since the key belongs to an entry that is foreign to the *books* table. For each table, an extra column may be created to store a unique ID. This is also called a key, but now a primary key.

Storing information like this, in a relational database, has proven to be extremely useful in the digital world. Information can quickly be obtained, and very few pieces of information, if any, needs to be stored twice. This is not only convenient in terms of file size, but it also means that if information needs to be updated, you only need to change it in one place and, based on that edited record, it is updated everywhere else. FileMaker allows you to point and click your way through setting up such a database, and by making attractive forms, filling that database with information can be both quick and somewhat fun. A big advantage of FileMaker is that you can create forms that work on iPhones and iPads. This allows for entering information on these handheld, touch screen devices, which then sync back to the main database on a computer.

What I had in mind doing was to load all the Evernote photos automatically in FileMaker, and then create different forms that each would only add a small piece of information. To start with, I wanted FileMaker to present me with a photo of a seal, and a small text box to type the index number visible on the seal. This would have been a task that I could do on my phone while waiting or traveling, and by accumulating all these small moments, I would have entered all the index numbers without truly having lost time over it. With the title pages, I wanted to do something similar, but with a twist. The first step was to get a photo of a title page on my screen, and then I would have to press and drag to create a rectangle around the title. After releasing, the photo would stay visible for another two seconds to allow the user to cancel; otherwise, the area of the rectangle would be stored as the area where the title is, and a new photo would instantly appear. Thus, it would be a matter of endless drawing of rectangles around titles. After that is done, a second step would be activated, in which only that part of the images as defined by the rectangles would appear on the screen, along with a text box, to type out the title. A similar strategy would be applied for each element on the title page. By chopping up the work into these small, menial tasks, I intended to catalog the collection in small, spare moments. The only problem was that to build that functionality in FileMaker would take a serious amount of time. Although I did have experience with FileMaker, I did not find the time to sit down and make it. I still think it is a good way to go about processing fieldwork, but it perhaps becomes more sensible when the corpus is bigger than a mere 1,500 entries and when there is a more immediate reason to get it done.

Instead, I fell back on a piece of software I had already been using for other parts of my research: Zotero. Zotero is a citation (or reference) manager, similar to EndNote and Mendeley, which syncs your references to its server. Zotero already provides that user-friendly interface I needed to type out the different details of the title pages. It has a function to export all the entries to a machine-readable format such as XML or JSON (more on this later). It does not work on phones or tablets since the interface of Zotero is designed to be used with a keyboard. After all, cataloging is generally a keyboard-reliant activity. I figured I could diminish that in FileMaker by creating custom inputs that would only require one or a few taps on the screen. Since Zotero or its third-party apps are not customizable, I had to change my workflow around the philosophy that cataloging should be done with a keyboard and in one go, collecting all metadata of one item before moving on to the next.

As a first step, I went through all the photos in Evernote where I had stored them, hand-typing the index number in the title of each note. This was a fairly painless job, since activating a note would instantly display the photo of the seal. This was more luck than wisdom, for if the photos of the seals came after the photos of the title pages, I would have had to scroll down on each note to reveal the index number. I figured it was worth the trouble of typing the index number in Evernote so that I could order the notes more easily and keep track of all of them better as I moved through the process of cataloging.

In the fall of 2017, I settled into a rhythm of working a little each night. Over two months, after some fifty hours,[5] I had pretty much completed typing out the catalog details. For some quick math, these fifty hours can fit neatly into two months if we assume an hour of work each day, spending about 2 minutes per item. Both estimates seem like a reasonable time. Little can be said about the use of Zotero since it is self-explanatory. The only odd thing is that the field to enter the location in the archive or the field to enter the call number is very far down. To reach them, I would have had to hit the Tab key a lot of times, which is both prone to mistakes and time consuming. Therefore, I ended up

[5] I know because I had the tv-show *The Office* on in the background and by the end of my manual data entry I had reached Season 8.

using the Language field to enter the index number, which is usually only two tabs down from the field for Year.

Having gone through my notes once, it was time to clean up the data I had entered. I ordered the entries on the Title field, which grouped together all entries with an Arabic title. I had typed these out in Arabic script, and now I added a transliteration in the Short title field. After ordering on the Place field, I consolidated place names, for example, changing all occurrences of 'Wien' to 'Vienna'. Ordering on the Language field, I could check the index numbers, making a note of those numbers that were missing. I identified 51 (about 3% of the collection), including those of which I had no photos, and those of which my photos were too blurry to make use of. I asked somebody with access to the library to investigate these numbers and send me photos of whatever could be found. As it turned out, after making a public catalog out of my data, I once again discovered some aspects that required cleaning up. For example, in Zotero, I had initially entered in the Year field whatever was on the title page, meaning that if there was a year from the Islamic calendar, I typed out that ending with an *h*, for *hijrī*. Only later did I notice that this will not allow ordering by year. So, instead, I converted all *hijrī* years to *mīlādī* years in the Year field and added to the 'abstract' field the original *hijrī* year. Having a graphically more attractive presentation of the catalog also enabled me to spot typographical errors and other mistakes more easily. After I had exhausted Zotero's functionality, I decided to export all the entries in the file format 'CSL JSON.' This stands for Citation Style Language JavaScript Object Notation.

To understand either part, CSL or JSON, let us first introduce a third term, XML, which stands for Extensible Markup Language. An XML-file is like a plain text ".txt" file, which you can open with any text editor. However, you are not supposed to simply type whatever you want, but you need to enter your information in a specific way for it to be a valid XML-file. This is because while an XML-file is easy to read for us human beings, by nature of the regular patterns of the specific way XML-files are supposed to be written, computers can interpret them easily too. This specific way is rather simple: every piece of information should be surrounded by tags. For example:

```
<example>some information</example>
```

The word 'example' is called the tag, and it is written between the angular brackets so that a computer can know this is the tag. As soon as a computer sees an < and an >, it will remember the word in between and look for the same word but this time in between </ and >. The slash, then, indicates a closing tag. Anything in between the opening and closing tag is the information related to the tag. Tags can exist within tags. For example, a description of a book can look like the following:

```
<book>
  <title>Among Digitized Manuscripts</title>
    <author>
      <firstName>L.W. Cornelis</firstName>
      <lastName>
        <surNameProper>Lit</surNameProper>
        <surNamePrefix>van</surNamePrefix>
      </lastName>
    </author>
  <publisher>Brill</publisher>
  <place>Leiden</place>
  <year>2020</year>
</book>
```

This format is probably quite easy to read for a person, although it is not a very attractive format. We can write code to have a computer go through it and place all the different elements in the right order using the right styling. For example, it can be printed to the screen as “Lit, L.W.C. van, *Among Digitized Manuscripts*, Leiden: Brill (2020).” The order, the addition of spaces, commas, and other punctuation marks, and the italics, are all done automatically—a great relief if you need to do this for hundreds of records or if you wish to change the styling later on.

XML does not impose any restrictions on what the tags should be. As long as all the tags close, it is valid. Whether the first tag reads *book* or *publication*, whether the nested tag in *lastName* reads *surNameProper* or *lastNameProper*, that is up to us. This makes XML usable for basically any situation in which some ordered data needs to be stored for computer manipulation. The drawback is, of course, that if I use *book* and you use *publication*, then a computer will not recognize them as the same. So, if somebody writes code that instructs the computer to take all the elements called *book*, it will not do anything if you have prepared a file where all these elements are tagged as *publication*. A standard that is accepted by everyone is needed, with rules that we all abide by, so that we can rely on the regularity of the rules to automatically extract and manipulate information from the XML-file.

Citation Style Language is such a standard. It is devised by companies behind three applications to manage references, Zotero, Papers, and Mendeley. CSL provides a way to combine all the different fields we used in Zotero into a list that can be read by any software that also uses CSL. For example, it will be very easy to export from Zotero and import into Mendeley.

What Zotero produces is, actually, not an XML-file, but a JSON-file. Opening a JSON-file in a text editor will demonstrate its similarity to XML. It has tags, that can be nested, containing information. The difference is that JSON does not need a closing tag and uses a shorter punctuation, making a JSON-file much smaller. Let us consider the above example, this time in JSON format:

```
{
  “title”: “Among Digitized Manuscripts”,
  “author”:
    {
      “firstName”: “L.W. Cornelis”,
      “lastName”:
      {
        “surNameProper”: “Lit”,
        “surNamePrefix”: “van”
      }
    },
  “publisher”: “Brill”,
  “place”: “Leiden”,
  “year”: “2020”
}
```

From a human point of view, it reads much more like a table, making it more readable. From a computer point of view, it is directly usable by one very popular programming language, JavaScript. Of course, other programming languages also know how to deal with it. For example, a dictionary in Python (see Chapter Seven) is very similar, and it takes little effort to load a JSON-file in Python as a dictionary. Since the name JSON reveals its affinity with JavaScript, by loading the above in a variable, say *book*, we have created an object *book*, and all the fields are its attributes. This means that if we ask JavaScript to print *book.publisher*, we get “Brill.” The question, then, of what to do with the catalog details of the Geyer collection in machine-readable plain text presents itself almost automatically: if we have made the catalog into a JavaScript Object, maybe we should make use of the web development technology to turn our data into a product that is easy to browse and search through; that is both pleasant to look at and pleasant to use. A printed publication did not seem the ideal solution from the start. Considering its small size and relative obscurity, it would hardly be commercially publishable, and printing it privately would mean the distribution would be poor and it would be costly, while the entire project had a budget of zero Euros. Delivering the catalog digitally, preferably online, was the best idea, and web development technology seemed the right tool for it.

2 Web Development: From My Computer to the World Wide Web

For students and scholars of the humanities, since more and more of our fieldwork will take place in this sphere, proficiency in web development is a desirable skill to have. Whether we want to use manuscripts or a catalog that a library makes available online, or whether we want to use Google Books, Facebook, Twitter, or Wikipedia to scrape information in order to map out discussions that take place on the internet, since these resources are built on web technologies, we also have to investigate them using web technologies. The digital world, after all, is for a rather large part built on web development technology.

The good news is that of all popular technologies, web development is among the easiest to learn, simply by the mere fact of the vast amount of teaching resources available. You can get very far, in fact, without paying a penny. However, the bad news is that web technologies are, at the time of writing, rapidly developing and changing, indicating that learning should take a two-pronged strategy. There are the basics that one simply has to know and will likely remain useful for many years to come, and there are the actual, fully-developed technologies that should be seen as electives and should be learned on a just-in-time basis.

For the technology of the catalog, I first considered some off-the-shelf products like DataTables and Omeka, which only require you to input raw data. I concluded that they had added unneeded features but lacked certain aspects that were an absolute requirement for our case. Rather than trying to hack them into the shape I wanted them to be, I decided to build it up from the ground, so that it would be exactly custom-fitted for our situation.

Because others have explained web technology much better and more detailed than I can, and because a lot of it undergoes rapid changes, I shall not go over the code in detail. Instead, I will give a more conceptual overview of how and why I put together the code that I used.

First of all: what does web development mean? This term encompasses all the technology required to develop things that are transmitted and received over a network. Usually, this network is the world wide web, and these things are websites. As the complexity of a website grows and offers more functionality to a user, we can better speak of a web application. For example, our catalog has all the functions of a catalog but is wrapped as a website, and users will have to reach it through a web browser.

The first basic division of web development is technologies that deal with the frontend and those that cover the backend. Frontend means what the user actually sees and can do in the browser, and backend is everything that goes on behind the scenes. Two similar (but not always the same) terms are client-side and server-side. Client-side refers to code that is downloaded and run on the computer of a user, while server-side code is run on a central server (the host of the website) and only gives the conclusion for the user to download and run. Server-side coding is largely synonymous with working with a database. A database is only necessary when the data of the web app is very large; it draws a lot of users, and/or its data requires very frequent updates or additions. Our catalog fits none of those descriptions. As a JSON-file, our catalog can be read by JavaScript and processed on the client-side.

What we want to develop, then, is a website that loads the JSON-file with all the catalog entries and displays them. We further want a simple search and a simple sort function, and we like the interface to be bilingual. It helps to sketch the layout of the website to understand what is needed.

It is useful to divide this type of development into four parts: one part of our code will govern the structure of our website, another will provide the content that goes into that structure, another piece of code will ensure how that content will be styled, and a final part will provide interactivity, not only between the user and the interface but also within the website itself. This division makes for flexible working. The structural level can ensure that there will be a Heading 1 title here or there, while the styling level can ensure that all those Heading 1's get, for example, a much bigger font size than regular text. On the level of content, we can define an English or German sentence for that heading while the interactivity can either program a button for the user to change the language or leave that functionality out and decide automatically which language to display.

The foundational programming languages for each of these levels are HTML for structure, JSON (or XML) for content, CSS for styling, and JavaScript for interactivity. They are each stored in a separate file, ending in either .html, .json, .css, or .js. We can have multiple files of each kind if a further distinction on each of these levels is required. For example, all the text for the interface and all the text for the catalog entries can be stored in two separate .json-files, for more clarity and flexibility. For example, if somebody asks for the catalog in JSON format, only one file needs to be given, about which nothing needs to be done, instead of having to open the one file containing both interface texts and catalog with a text editor, cutting out all the parts related to the interface, saving it as a separate file, and only then sending it.

Because of the flexible structure with which we build this catalog, you will notice quickly that we are not so much developing a catalog for Rudolf Geyer, but a generic catalog into which you can pour any JSON that Zotero creates, with only a few adjustments that need to be made. The result would be a functioning, minimalist catalog appearing on the screen. I shall go over each part separately but since the most is happening in the JavaScript that part will take up the most space.

3 Structure: HTML

HyperText Markup Language was originally conceived to fulfill all the four roles that I just separated. It is the foundational language with which to create webpages that a browser, the application by which you surf the internet, can read and display correctly. For example, if some text is placed within b-tags, you would not see in the web browser <b>some text</b>, but you would see **some text**: the browser reads the b-tags and knows that that is an instruction to display the text in between in bold. Your web browser, then, knows how to read and display an HTML-file, but would not know what to do with any of the other files. So, in this sense, the HTML-file is the gateway to the rest of your online product (whether it be a catalog or something else). In fact, currently, the only thing we need is an *index.html*, which is traditionally the name of a website's landing page.

If you look over the code of the HTML-file, you will notice that it basically serves two functions. First, the *head*-tag is a shell that opens all the other JSON, CSS, and JS files. Second, within the *body*-tag, the HTML-file dictates a structure of where the different elements of our catalog should go.

First, a header is defined, which is divided into three parts. The header, as a whole, stands out for its different background color. On the top-left of the page, we want a logo of Stift Florian that links to the main website of the monastery. In the middle, we want a title that should look big and prominent, so I made it a Heading 1 title with the *h1*-tags. On the right, we want to have a button to change the language and a button to go to my own website (as a small nod to me being the creator of the catalog). The button to change the language should be a flag representing the language into which the user can change it. If the interface is currently in German (which it will be upon first loading the page, as

```
<!-- Copyright L.W. Cornelis van Lit, Please see https://
github.com/LWCvL/RudolfGeyerCatalog for details. -->
<!DOCTYPE html>
<html lang=“en”>

<head>
  <meta charset=“UTF-8”>
  <meta name=“viewport” content=“width=device-width,
initial-scale=1.0”>
  <meta http-equiv=“X-UA-Compatible” content=“ie=edge”>

  <!-- Dependent on MicroModal, Bootstrap, Popper and
TippyTip -->
  <script src=“https://unpkg.com/micromodal/dist/micromodal.
min.js”></script>
  <link rel=“stylesheet” href=“https://maxcdn.bootstrapcdn.
com/bootstrap/4.1.1/css/bootstrap.min.css”>
  <script src=“https://maxcdn.bootstrapcdn.com/
bootstrap/4.1.1/js/bootstrap.min.js”></script>
  <script src=“https://unpkg.com/popper.js@1”></script>
  <script src=“https://unpkg.com/tippy.js@4”></script>

  <!-- Loading interface texts and catalog texts -->
  <script src=“textsGeyerCatalog.json.js”></script>
  <script src=“textsInterface.json.js”></script>

  <!-- Loading styling rules overriding standard Bootstrap -->
  <link rel=“stylesheet” href=“extrastyle.css”>

  <title>Catalogue of Rudolf Geyer</title>
</head>

<body>
  <!-- Header with link to St. Florian, a title, link to
switch language, and link to Digital Orientalist -->
  <header>
    <div class=“container-fluid”>
      <div class=“row”>
        <div class=“col-2 headerdiv left”>
```

I instructed in JavaScript), the flag will be of the United Kingdom, representing English. If the interface is in English, it will be a German flag. The button to go to my website should be the logo of The Digital Orientalist, but it will be quite small. It should change color when the user hovers their mouse over it—to emphasize it and indicate that it is a button.

As you may notice, all the images are SVG-images. These are images that do not store color values pixel per pixel, but rather store the coordinates of shapes and their colors in an XML-like manner, which browsers nowadays know how to turn into images. They are essentially connect-the-dots puzzles that the computer of the user will solve on the spot. This means that these images are vector-based and will look sharp no matter how small or big you make them. It also means you can edit them in the text editor where you edit your code (see the Productivity section below), for example, to change the aspect ratio or the color.

Below the header comes an introduction section. It is divided into ¾ text and ¼ image. The image is Rudolf Geyer's Ex Libris sticker, which we will discuss further in the CSS section. The text consists of a heading and an actual introduction text. As you may notice by now, the HTML-file does not have the actual title or text, as these are stored in a JSON-file in two languages, and JavaScript will fill these different blocks with the correct texts. It can do so by calling the different elements in the HTML-file by their ID-name, such as *welcomeHeader* and *welcomeBody*. The frequent mentioning of *class* is to give the CSS-file a sign that this or that part needs to be formatted in a certain way.

Below the welcome text, a text box and a button needs to appear to give the user the opportunity to search. Under the search function, a horizontal line should be drawn, indicated by <*hr*>, to separate the interface from the actual catalog.

At the very top of the catalog, a title should appear. As we will see, this title changes frequently, for example, notifying the user of the number of search results. Next to the title, preferably on the same line, three buttons should appear on the far-right to sort the displayed entries (the entire catalog or just the search results) according to the title, author, or year. I did not deem necessary anything more than this. If a place name or something else is important, one can simply search for it.

After the title, an unassuming <*div class='container' id="catalogGeyer"*></*div*>, to be filled by a JavaScript function, defines the entire catalog. Similarly, the *popupInfo* plays an important role later and simply needs to be declared here as part of the structure. With it, we can program a pop-up to show more details of a catalog record. This is necessary since we just defined a lightweight,

```
          <a href=“http://www.stift-st-florian.at/en/home.
html” class=“btn” id=“stiftFlorianTip” title=“”>
            <img src=“iconStFlorian.svg” height=“30”
alt=“Stift Florian”>
          </a>
        </div>
        <div class=“col headerdiv middle”>
          <h1 class=“header” id=“headline”></h1>
        </div>
        <div class=“col-2 headerdiv right”>
          <div class=“flags btn” id=“flagTip”>
            <img src=“iconflaggerman.svg” class=“flags”
height=“16” width=“32”>
          </div>
          <div class=“flags btn DO”>
            <img src=“iconDigOr.svg” class=“flags”
height=“16” id=“digitalOrientalistTip” onclick=“window.
open(‘http://www.digitalorientalist.com/’)”
              onmouseover=“this.src=‘iconDigOrOrange.svg’”
onmouseout=“this.src=‘iconDigOr.svg’” />
          </div>
        </div>
      </div>
    </div>
  </header>

  <!-- Space for welcome text -->
  <p></p>
  <div class=“container”>
    <div class=“row”>
      <div class=“col-9”>
        <h2 class=“headline” id=“welcomeHeader”></h2>
        <p id=“welcomeBody”>
          Loading...
        </p>
      </div>
      <div class=“col-3 geyerImage”>
      </div>
    </div>
    <!-- Search bar -->
    <p class=“font-weight-bold”>
      <div class=“input-group”>
        <input type=“text” id=“searchTip” class=“form-control”
title=“” placeholder=“”>
        <span class=“input-group-btn”>
          <button class=“btn btn-blue” type=“button”
id=“searchButton”></button>
        </span>
      </div>
    </p>
  </div>

  <!-- Dividing line interface/catalog -->
  <hr>

  <!-- Title and buttons for ordering catalog -->
  <div class=“container”>
    <div class=“row”>
      <div class=“col-9” id=“titleCat”>
        <h2 class=“headline” id=“titleSearch”>
        </h2>
      </div>
      <div class=“col-3” id=“buttonsCat”>
        <span id=“sortBy”></span>
        <div class=“btn-group btn-group-toggle” id=“ordering”
data-toggle=“buttons”>
          <label class=“btn btn-outline-secondary”>
            <input type=“radio” name=“options” id=“sortTitle”
autocomplete=“off”>
            <span id=“sortTitleCaption”></span>
          </label>
          <label class=“btn btn-outline-secondary”>
            <input type=“radio” name=“options”
id=“sortAuthor” autocomplete=“off”>
            <span id=“sortAuthorCaption”></span>
          </label>
          <label class=“btn btn-outline-secondary”>
```

```
            <input type=“radio” name=“options” id=“sortYear”
autocomplete=“off”>
            <span id=“sortYearCaption”></span>
          </label>
        </div>
      </div>
    </div>
  </div>

  <!-- Actual catalog -->
  <div class=‘container’ id=“catalogGeyer”></div>

  <!-- Optional pop-up for extra details of an entry -->
  <div id=“popupInfo”></div>

  <!-- Footer with link to Stift Florian and link to Digital
Orientalist -->
  <footer>
    <div class=“container-fluid”>
      <div class=“row”>
        <p></p>
      </div>
      <div class=“row”>
        <div class=“col-2 footer left”>
          <a href=“http://www.stift-st-florian.at/en/home.
html” class=“btn” title=“”>
            <img src=“iconStFlorian.svg” height=“30”
alt=“Stift Florian” srcset=“”>
          </a>
        </div>
        <div class=“col footer middle”>
        </div>
        <div class=“col-2 footer right”>
          <a href=“http://www.digitalorientalist.com/”
class=“btn” title=“”>
            <img src=“iconDigOr.svg” height=“20”
alt=“The Digital Orientalist” srcset=“” onmouseover=“this.
src=‘iconDigOrOrange.svg’” onmouseout=“this.src=‘iconDigOr.svg’”>
```

```
          </a>
        </div>
      </div>
    </div>
  </footer>

  <!-- Loading of functionality -->
  <script src=“functions.js”></script>

</body>

</html>
```

compact catalog that would in the first instance only show the title, author (and editor), place, and year. A fuller record can then be shown when the user clicks on the entry.

We finally arrive at the footer of the page, which looks a lot like the header but without the ‘change language’ button. We close with a reference to a JavaScript file that actually is the motor of all our interactivity. It needs to be placed here at the end, for otherwise, it will not function properly. This is because it can only interact with things that are defined before it, so if it is placed at the top of the HTML-file, it would already start running and would want to try to refer to things that are not yet defined, resulting in nothing.

4 Content: JSON

If you have gone through my repository, you will have noticed that there are two JSON-files. In this case, they actually end in the extension .JS, as will be explained later on. One handles the interface (for both languages), the other contains the catalog. I structured the interface-file in a simple manner. At the top level, I defined the keys (the string on the left of the colon) as the same as the ID that the elements in the HTML-file have and to which the different texts belong. From a human-readability point of view, this should make it quite clear where each of these texts goes in the final website. In terms of the vertical structure, the order in which the keys are defined, it seemed to make the most sense to be as faithful to the structure of the website. Thus, header texts are first defined, and the texts for the modal, the pop-up showing the full record details, appear last.

```
var texts =
  {
    “headline”:
      {
        “english”: “Catalogue of the Rudolf Geyer Collection”,
        “german”: “Katalog von dem Nachlass Rudolf Geyer”
      }
    ,
    “stiftFlorianTip”:
      {
        “english”: “Go to Stift Florian’s Website”,
        “german”: “Gehe zur Website von Stift Florian”
      }
    ,
    “digitalOrientalistTip”:
      {
        “english”: “Made by L.W.C. van Lit”,
        “german”: “hergestellt von Dr. L.W.C. van Lit”
      }
    ,
    “flagTip”:
      {
        “english”: “Switch to English”,
        “german”: “Auf deutsch umstellen”
      }
    ,
    “welcomeHeader”:
      {
        “english”: “Welcome”,
        “german”: “Herzlich willkommen”
      }
    ,

var catData = [
  {
    “id”: “http://zotero.org/users/170941/items/8JKDXVME”,
    “type”: “book”,
    “title”: “كتاب شعراء النصرانية”,
    “publisher”: “Maṭbaʿat al-ābāʾ al-mursalīn al-yasūʿiyyīn”,
    “publisher-place”: “Beirut”,
    “event-place”: “Beirut”,
    “abstract”: “al-juzz al-awwal fī shuʿarāʾ al-jāhiliyya”,
    “shortTitle”: “Kitāb shuʿarāʾ al-naṣrāniyya”,
    “language”: “XVI A 1”,
    “editor”: [
      {
        “family”: “Sheikhu”,
        “given”: “Louis”
      }
    ],
    “issued”: {
      “date-parts”: [
        [
          “1890”
        ]
      ]
    }
  },
  {
    “id”: “http://zotero.org/users/170941/items/DWFZQGTU”,
    “type”: “book”,
    “title”: “The Poems of Ṭufail ibn ʿAuf al-Ghanawī and aṭ-
Ṭirimmāḥ ibn Ḥakīm aṭ-Ṭāʾyī”,
    “publisher”: “Luzac & Co.”,
    “publisher-place”: “London”,
    “event-place”: “London”,
    “abstract”: “Arabic Text Edited and Translated. \nPrinted
for the Trustees of the ,,E. J. W. Gibb Memorial’’”,
    “language”: “XVI A 2”,
    “editor”: [
      {
        “family”: “Krenkow”,
        “given”: “F.”
      }
    ],
    “issued”: {
      “date-parts”: [
        [
          “1927”
        ]
      ]
    }
  },
```

Each of these key:value pairs contain another object as the value, with the first key being 'english' and the second being 'german.' Their values are semantically the same, but they store the required text in two languages. This structure will allow for easy deleting, swapping, or adding of languages. If we would want to add an interface in Arabic, we would only have to go through this JSON-file and write a comma after each 'german' value, hit Enter, type "arabic": followed by the text in Arabic in between quotation marks. This is, I think, a better structure than, for example, having at the top level the keys of each language followed by an object containing all the interface elements, because if we would want to adapt some part of the interface, for example, the text of the introduction, we would have to scroll to different places in this file to change the text for each language. In this case, we only need to go to one place and instantly see the same text in different languages, making it easier to accurately change the text for each language.

For the JSON-file generated by Zotero, the only change that I made was to put var catData = at the very beginning of the file and change the extension to .js. This has a technical reason. By using it as a JavaScript file, the loading can simply be done by an HTML-command starting with <script src="…, and then the data is loaded. To make that a valid process, the data still needs to be a JavaScript Object, meaning that it needs a variable name to be stored in. In our case, this variable is called catData (short for catalog data). Once this file is loaded, we can use the catalog data by invoking catData, as we will do in the JavaScript part of this web app. The JSON-file that governs the interface texts was treated similarly, meaning it was in the end changed to a .JS extension, a variable name was included at the beginning, and the file was loaded using the HTML-command. If we had kept the files with a .json-extension, we would have to have used a special JavaScript command called XMLHttpRequest to load the files to the user's computer memory. This command, however, only runs when it is on a server. JavaScript cannot access local files, and this is for security reasons. Understandable as this may be, the use of the XMLHttpRequest command is not a good alternative in our case. This would make it impossible to download all the files belonging to the catalog, opening it on a computer, and making use of it since the files would run on a server. This is a moment where we have to be creative with the available technology and not dive headfirst into making the XMLHttpRequest command work. Rather, we should reconsider our options. The first step, in this case, is to reflect on the typical use case worldwide for JSON. The dominant usage is for sending or receiving data between machines. For example, when we let JavaScript talk to Wikipedia or Weather Underground or any other service (known as an API), the response will be in the shape of a JSON. We may notice that in our case we own the data

ourselves, and it does not change over time (or only rarely do so), so we do not need the same kind of dynamic set up to collect our catalog data. In that case, we might as well change them into JS files and add the extra line at the beginning to give the object a name. This does not change their being and function; we can still treat them as JSON-files, and hence, it will be clearer if we include '.json' in the filename.

It seemed a good idea to use the JSON-file created by Zotero with the smallest amount of tinkering so that we could create a different catalog in a pinch by exporting a different set of entries from Zotero. There are, unfortunately, some drawbacks to this, owing to the particular structure of JSON-files created by Zotero. The chief drawbacks are elements like *ID* and *event-place,* which are pretty much useless for our purpose, but it may also be noted that these files have an overly complicated way of storing names and dates. Once again, we need to reassess the possibilities of the technology to find our best strategy to deal with these drawbacks. One solution is to fiddle around with the JSON-file that Zotero produced and shape it into a much simpler form, a form exactly as we want it. We can do this by using the search and replace function, with regular expressions to capture the part in every entry that we want to be deleted or changed.

In the end, I decided to not make any alterations to the JSON-file but instead make the JavaScript code that extracts information from the catalog more complicated. This way, the catalog is more faithful to the data entry in Zotero, and it will be easier to implement it for another collection.

5 Style: CSS

CSS stands for Cascading Style Sheets. It is one or more sheets (files) that end in .css and which should be called at the beginning in the HTML-file with a <link rel="stylesheet" href="pathAndNameOfStylesheet.css"> command. Such a CSS-file is merely a list of instructions regarding how particular elements should look like. Proper styling is especially necessary as there is an ever larger variety of screen sizes and browsers that users of your product will use. Taking this into account is using what is called 'responsive web design,' which will ensure that your website will not shrivel to an unreadably small size because someone uses a four-year-old smartphone, but instead, it will respond by reshuffling and resizing the elements and the texts in them. Styling is also necessary to make the website actually attractive. A catalog entry stored in a JSON-file is sort of readable by an individual, but once it is presented in a table with ample white space around it, lined out neatly, with good background colors and a font and font style that fit the context, it will be much more readable and enjoyable. In fact, as more and more websites do look good, users are simply starting to expect a certain level of sharp and attractive design.

```
headline {
  background-image: linear-gradient(to bottom right, rgb(167,
72, 61), rgb(253, 192, 47));
  color: transparent;
  -webkit-background-clip: text;
  background-clip: text;
  }
.geyerImage {
  background:
url("GeyerExLibris1.jpg") no-repeat;
   background-size: contain;
   background-repeat: no-repeat;
   width: 100%;
   height: 100%;
   padding-top: 27%; /* (img-height / img-width * container-
width) */
   -webkit-background-size: contain;
   -webkit-transition: all 1s ease-in-out;
   -webkit-border-radius:  10px;
  }
```

Whereas the structure of our website in HTML really is up to us, for CSS, we can and should use already existing resources that are capable of instantly making our site reasonably usable and attractive. Styling controls the look and feel of every aspect of the website, so starting from scratch would be a tedious job, especially since many things like how a button looks and behaves should be sort of similar regardless of the online resource you are building. The responsive part of CSS, too, is equal no matter which project you are working on. Two templates that are currently popular are Bootstrap and Materialize. The first was created internally at Twitter and subsequently released as open source. The second was developed later by students of Carnegie Mellon University. I have used Bootstrap for my catalog. The changes that I wished to make have been put in an additional CSS-sheet, and I have instructed the HTML-file to load this additional file after loading the Bootstrap CSS-file. The 'cascading' in Cascading Style Sheets means that if there are two rules about the same thing, it is the last rule that is actually obeyed. Thus, by loading a custom made CSS-file after Bootstrap, we can use Bootstrap as a foundation and make some over-ruling changes to it, defined in the other file. Knowing what exactly can be controlled and which commands to use is a matter of searching the internet, specifically websites like StackOverflow and the documentation for CSS and for Bootstrap. For beginners, it will be beneficial to watch a video or attend a workshop. In most cases, it is only when the need arises that you should look into exactly how to do something.

Perhaps the most noteworthy change is the definition of *headline* and of *geyer*. Headline is used for different headings, such as the title of the welcome text and the title of a holding in the pop-up. In my extra CSS, I made it so that the text of those headings are colored in a gradient. Done well, this can really pop without being overwhelming. Geyer is only used on one element in my HTML-structure, concerning the image that is supposed to appear next to the welcome text. Even though JavaScript is officially in charge of interactivity, you can see that I actually added some interactivity right here in the CSS-file with the transition command. The command needs to be repeated for several different browsers for it to work for all users. CSS also allows to style elements specifically when a user hovers over the element with the mouse pointer. In this case, a similar transition command is used but on a different image. It results in the following effect: when the user hovers over the image, one Ex Libris sticker image appears to morph into another Ex Libris sticker. Note the *width, height*, and *padding-top* instructions, which ensure that the image takes up the entire width of the defined area, which will differ from screen to screen, while still showing the entire height of the image proportionally to the width so that it always retains its correct height to width ratio. Without the padding-top instruction, the image would likely be cut off at some point.

6 Interactivity: JavaScript

In about five-hundred lines of code, the catalog comes to life. These lines are divided over 12 functions which, in some cases, require human interaction to be triggered and, in other cases, are auxiliary to other functions. If you open the JavaScript file in a good code editor and select *Fold All*, the editor will reduce all groups of code that belong together to a single line. In our case, this means that you will instantly see an overview of the division of functions and their triggers.

```
// Copyright L.W. Cornelis van Lit. For details see https://
github.com/LWCvL/RudolfGeyerCatalog
// (f) = function, (e) = event

// 1. Initializing webapp
window.onload = function() {
  // setting initial variables
  searchData = catData;
  shadowCatalog = {};
  accentMap = {
á: “a”, à: “a”, ā: “a”, â: “a”, ä: “ae”, æ: “ae”, é: “e”, è:
“e”, ē: “e”, ë: “e”, í: “i”, ì: “i”, ī: “i”, î: “i”, y: “i”,
ó: “o”, ò: “o”, ō: “o”, ô: “o”, ö: “oe”, ú: “u”, ù: “u”, ū:
“u”, û: “u”, ü: “ue”, ṭ: “t”, ṣ: “s”, ḍ: “d”, ḥ: “h”, ẓ: “z”,
ġ: “gh”, š: “sh”, ğ: “j”, ḳ: “q”, č: “ch”, ‘: “’”, ’: “’”,
“.”: “ ”, “,”: “ ”, “:”: “ ”, “;”: “ ”, “?”: “ ”, “!”: “ ”
  };

  // Initial building of shadow catalog and interface
  Reducing_Catalogue(catData);
  Switch_Language(“german”);
};

// 2. Creating a shadow-catalog with simplified entries to
easier match a search term in the search function
// Called by (e)window.onload
function Reducing_Catalogue(catalog) {
  // Looping through all catalogue entries
  for (var y = 0; y < catalog.length; y++) {
    // Adding all fields for each entry in one string.
    entryY = “”;
    entryY += “ ” + catalog[y].title;
    entryY += “ ” + catalog[y].shortTitle;
    entryY += “ ” + catalog[y].publisher;
    entryY += “ ” + catalog[y][“publisher-place”];
    entryY += “ ” + Print_Names(catalog[y].author);
    entryY += “ ” + Print_Names(catalog[y].editor);
    entryY += “ ” + Print_Names(catalog[y].translator);
    entryY += “ ” + catalog[y].abstract;
    entryY += “ ” + catalog[y].language;
```

The order in which the functions are written here is based on their logical order, but I did not think all of this out at first, and I did not write these twelve functions neatly in this order. As is often the case, you simply start coding functionality that seems to be of the most immediate need from a usability point of view. By constantly testing the result in a browser, you will see how the code behaves. You will quickly run into issues in how programming languages require you to think and the limitations that JavaScript specifically put on you. At each step, it is good to consider if the functionality you are now thinking of writing is essential, and if the way you want to build that functionality is efficient.

Functions four and five form the core of this code, and to understand the entirety of it, it will be better to start there. The task of function four is to fill the interface with texts. Moreover, when the flag icon at the top-right of the website is clicked, it needs to change the texts to the next language. To keep the structure and content of the website strictly separated, I only declared an empty structure in the HTML-file, which means that when the user visits the website, this function needs to be called to populate all the fields. The function does not make many assumptions. For example, it does not specify the different fields by name but loads the name of those fields dynamically from the interface-JSON. This means we could add or delete certain interface elements by changing the HTML-file and the JSON-file, while leaving this code intact. Similarly, the number or kind of languages there are is not specified. Adding another language to the interface-JSON will be of no problem as long as one does not forget to also create a flag icon in SVG-format for that language. The language the interface is in right now is given to the function, and from there, the next language is established, both the index of the next language and the name. This index is used to access the correct text for all the fields, which is done by a simple *for*-loop, going through all the available elements. Within this loop, some specificity was inescapable, as accessing the text property of all fields could not be done with one command. Some require the command *placeholder*, others *innerHTML*, and others *title*.

Function five, *Render_Table*, creates the catalog underneath the interface. It takes a couple of parameters. First, the catalog to be rendered, in JSON format. Then a specification regarding the type of heading the catalog should have. If you look into the interface-JSON, you will see several different headings such as *beginTitleCatalog* and *noResultsTitleCatalog*. This function takes one of them to give a more specific feel to the rendered catalog. Next, the function takes in the language. Lastly, it takes the number of entries to be rendered, which could have been calculated within the function, but it seemed more readable to parse it into the function. The logic of the function is rather simple: it will

```
    // If field does not exist, ‘undefined’ will be pushed.
All ‘undefined’s are now deleted.
    var definedY = entryY.replace(/undefined/g, “”);
    // Entire string is stripped of transliteration marks,
punctuation, capitals, and double spaces, and pushed into the
shadow catalogue
    shadowCatalog[y] = Simplify_Term(definedY);
  }
}

// 3. Returning simplification of transliteration signs,
punctuation, double spaces, and capitals
// Called by (f)Reducing_Catalogue and (f)Search_Catalog
function Simplify_Term(inputWord) {
  if (!inputWord) {
    return “”;
  }
  inputLower = inputWord.toLowerCase();
  var outputWord = “”;
  // Loop through every letter of a word
  for (var i = 0; i < inputLower.length; i++) {
    outputWord += accentMap[inputLower.charAt(i)] ||
inputLower.charAt(i);
  }
  returnedNoSpaces = outputWord.replace(/\s{2,}/g, “ ”);
  return returnedNoSpaces;
}

// 4. Switching interface to different language (currently
German-English)
// Called by (e)window.onload and (e)flagTip.onclick
function Switch_Language(language) {
  // Array of words that are both keys to the texts-Object
and elementIDs of the webapp
  toBeTranslated = Object.keys(texts);
  // Array which is agnostic as to how many/which languages
there are
  languages = Object.keys(texts[toBeTranslated[0]]);
  // index number of current language
  lanIndex = languages.indexOf(language);
```

create a variable containing a string that represents the HTML-code to display the catalog, for which the *table*-tag is used, and once it has filled that variable with all entries, it will fill the *div* in the HTML-file called *catalogGeyer* with that variable. The last thing it does is set the heading for the catalog.

It seemed cleaner and more flexible to take out the code for rendering one catalog entry as one table item and put it into a separate function called *Render_Table_Entry*. At an early stage, I realized it might be the case that a search function would require the code of this *Render_Table_Entry* function separate from the *Render_Table* function, in which case there are logical grounds to separate the code out as a function. Writing the same code twice or more is bad practice. Not only does it make your code longer than necessary, but it also means that if you want to change something in it, you would also need to change it in all those places. In the end, it turned out I would only call this function from within the *Render_Table* function, so separating the two functions has no logical reasons, but it is still cleaner and more readable. First, the *Render_Table_Entry* function declares a few variables that are filled with the correct texts in the correct language. The only reason to declare those variables is to make the rest of the code more readable. The HTML for the table-entry is generated piece by piece. First, a 'more information' button is generated, which triggers a pop-over that we will discuss later. Then, a variety of different things are included, such as title, author, place, and date. Since some of these things are missing from the catalog entry, they should only be included if they are there. Since the name fields can include many names—multiple authors or multiple editors, for example—one unified function to print these correctly is called, which we will discuss next. For the date and place, we cannot use a *hasOwnProperty*-function since the data structure that Zotero produces is a bit more complicated than the other ones, and JavaScript simply cannot handle this function for that data structure. Writing merely the first part of the name of that data structure will be enough to check for its existence. We could do likewise for the previous ones, the names, but the actual function *hasOwnProperty* seems more readable. Also notice that if the key in a JSON has a dash in it, like *publisher-place*, it cannot be accessed with a dot notation like *data*[*i*].*publisher-place*, but we need to use square brackets and quotes like *data*[*i*][*"publisher-place"*].

Print_Names, function number eight, takes a reference to a catalog entry and returns the names associated with that entry with appropriate formatting. The first thing the function does is to ensure the existence of names. If that is the case, a *for*-loop iterates over all the names and for each name checks if there is a first name, a last name, and a connecting particle (such as the German 'von' or the French 'de'). Since Zotero stores these values in separate

```
  if (lanIndex == languages.length - 1) {
    nextLanIndex = 0;
  } else {
    nextLanIndex = lanIndex + 1;
  }
  nextLanguage = languages[nextLanIndex];

  for (key in toBeTranslated) {
    // Filling different elements of webapp by their ID. Some
need different ways of accessing their content.
    if (toBeTranslated[key].includes(“Tip”)) {
      document.getElementById(toBeTranslated[key]).title =
Object.values(
        texts[toBeTranslated[key]]
      )[lanIndex];
    } else if (toBeTranslated[key] == “searchBox”) {
      document.getElementById(“searchTip”).placeholder =
Object.values(
        texts[toBeTranslated[key]]
      )[lanIndex];
    } else if (toBeTranslated[key].includes(“Catalog”)) {
      null;
    } else {
      document.getElementById(toBeTranslated[key]).innerHTML
= Object.values(
        texts[toBeTranslated[key]]
      )[lanIndex];
    }
    // Setting button for switching language to the next one
    document.getElementById(“flagTip”).innerHTML =
      ‘<img src=“iconflag’ +
      nextLanguage +
      ‘.svg” height=“16” width=“32” class=“flags”
id=“flagTip” )”>’;
    document.getElementById(“flagTip”).title = Object.
values(texts.flagTip)[
      nextLanIndex
    ];
  }
```

fields, we need to stitch them together. All names are separated by a comma (and a space, of course), and the last name is closed with a period.

Function seven, *Popup_More_Info*, handles the event that a user clicks the circle with an 'i' in it to get more information about a catalog entry. The pop-up relies on what is called a 'modal,' a nice looking overlaying sheet which can be closed by either clicking a close button or by clicking anywhere outside of it. We rely on a third-party library. This is because it takes remarkably smart code to get a really good modal, and if somebody else already offers it for free, we are better off using that. In our case, we use MicroModal. In the HTML-file you may have noticed that we load a JavaScript file called micromodal.min.js; this is what makes it possible to refer to the MicroModal functionality in this *Popup_More_Info*-function. If a better modal library would come along, I could swap out this one for the other relatively easily. In fact, I already did so. Before I knew of MicroModal I used jQuery. jQuery is a multi-function library that has become outdated if used as a framework and it seemed that for a modal, MicroModal performed better.[6] Without getting too bogged down with details about libraries and frameworks, let us consider what this function does because if it works, it is probably good enough for our use case. In the beginning, several variables are defined whose sole purpose is to make the following code more readable. They consist of all the particulars of a catalog entry. Just like creating the table for the catalog was done by filling a variable with the right HTML-encoding and then filling a div in the HTML-file with that variable, so we do the same for the pop-up. To get the modal to work, we need to add a couple of different divs. They can be copied and pasted from an example on the internet. Then, one after the other, the particular details of a catalog entry are considered. We perform a check to see if the entry has a certain detail, and if it is there, it is added to the variable. Notice the use of markup tags like <i> for italics and reference to CSS elements such as class= "text-success" for the abstract.

With these functions, we now have a visible catalog. Of course, we want to add a search and a sort functionality. There are a couple of simple commands we can use to get that working. For the search function, we can use the JavaScript command *.includes*,[7] which will return *True* if the term is part of the catalog entry (it really is that easy). For sorting the entries, we can use *.localeCompare* for strings (such as titles and author names) and *.sort* for numbers

[6] You can see the changes by looking through the commit history of my GitHub repository for this code.

[7] I first considered filtering the generated table, but this caused a number of issues which I do not think are worth going into.

```
  // Catalogue needs to be re-rendered
  if (document.getElementById(“searchTip”).value) {
    Search_Catalog(document.getElementById(“searchTip”).value);
  } else {
    Render_Table(catData, “beginTitleCatalog”, lanIndex);
  }

  // Fancy mouse-over tool-tip only activitated when not on
mobile. We simply delete previous instances and create new ones
  [...document.querySelectorAll(“*”)].forEach(node => {
    if (node._tippy) {
      node._tippy.destroy();
    }
  });
  if (
    /Android|webOS|iPhone|iPad|iPod|BlackBerry/i.
test(navigator.userAgent) ==
    false
  ) {
    tippy.setDefaults({
      animation: “perspective”,
      arrow: “true”,
      size: “large”
    });
    tippy(“#flagTip”, {
      content: Object.values(texts.flagTip)[nextLanIndex]
    });
    document.getElementById(“flagTip”).
removeAttribute(“title”);
    tippy(“#digitalOrientalistTip”, {
      content: Object.values(texts.digitalOrientalistTip)
[lanIndex]
    });
    document.getElementById(“digitalOrientalistTip”).
removeAttribute(“title”);
    tippy(“#stiftFlorianTip”, {
      content: Object.values(texts.stiftFlorianTip)[lanIndex]
    });
    document.getElementById(“stiftFlorianTip”).
removeAttribute(“title”);
    tippy(“#searchTip”, {
```

(such as the publication year). What requires some extra work is to get this functioning amidst the messy reality of our catalog. For example, when we search for "tufail," we want the *.includes*-command to return *True* for a catalog entry which includes "Ṭufayl." While our human eyes and brains instantly see the equivalence of tufail and Ṭufayl, a computer strictly compares the character set of the two terms, concluding that they are not the same. Similarly, when sorting entries by year, if a publication does not have the Gregorian year 1871 printed on the cover but instead the Hijrī year of 1288, it should not be sorted as though it comes from the Gregorian year 1288, but it should be sorted amidst books from 1871. While some of these things are better taken care of within JavaScript, some are not. For the year issue, as mentioned before, it was easier to normalize the data entry. That is to say, I went back to Zotero and calculated the Gregorian year for every date only given in the Hijrī calendar. I then added the Hijrī date to the 'abstract' field. Finally, I exported all the catalog entries again to obtain an updated JSON-file.

For the issue of typographically different but semantically equivalent strings, we can use function three, *Simplify_Term*. This function takes any string, changes all upper case letters into lower case, and then applies a series of reductions to it that are specifically designed for the kind of terms and letters used in transliterating Islamic languages (in various transliteration schemes). This schema needs to be defined somewhere. It could have been stored in the interface-json, but I decided to store it in the JavaScript file. It is called *accentMap* and sits in the first function, which is executed as soon as somebody enters the website. Here, we can see how an *ā* is defined to be reduced to an *a*, and an *š* becomes an *sh*, and so forth. Thus, *Šāhnāma* becomes *shahnama*. The function *Simplify_Term* goes through every letter of the string it is given and performs the reduction on it. Then, it deletes any extra spaces after which it returns the resulting string.

With the function *Simplify_Term* discussed, we can look at function two, *Reducing_Catalogue*. After experimentation, I found out that the search function probably worked best if we created another version of the catalog that would only contain the entries in reduced form. *catData* is the object that contains the catalog, and *searchData* will now become the object with the reduced catalog, what I call a shadow catalog. At first, I used the simple commands *.values*, which would give you all the values of an entry, such as title, place, and year; and I used *.join*, which takes all those values and stores them together in one long string. Elegant as this solution was, it was not performing well because the date and names are stored in a more complicated form than this approach allows. I finally settled on filling a variable with individual values of an entry, making profitable use of the function *Print_Names*. Since I did not perform

```
      content: Object.values(texts.searchTip)[lanIndex]
    });
    document.getElementById(“searchTip”).
removeAttribute(“title”);
  }
}
document.getElementById(“flagTip”).onclick = function() {
  Switch_Language(nextLanguage);
};

// 5. Rendering the table of catalogue entries
// Called by (f)Switch_Language and (f)Search_Catalog
function Render_Table(data, heading, language, numberEntries) {
  document.getElementById(“catalogGeyer”).innerHTML = “”;
  htmlTableCat = ‘<table id=“tableCatalog” class=“table
table-striped”><tbody>’;
  for (i = 0; i < data.length; i++) {
    Render_Table_Entry(data, i, language);
  }

  htmlTableCat += “</tbody></table>”;

  document.getElementById(“catalogGeyer”).innerHTML =
htmlTableCat;
  if (typeof numberEntries !== “undefined”) {
    document.getElementById(“titleSearch”).innerHTML =
      Object.values(texts[heading])[language] + “ ” +
numberEntries;
  } else {
    document.getElementById(“titleSearch”).innerHTML = Object.
values(
      texts[heading]
    )[language];
  }
}

// 6. Rendering one entry of the table
// Called by (f)Render_Table
function Render_Table_Entry(data, i, language) {
  var authorLan = Object.values(texts.authorCatalog)[language];
```

if-then checks to see if a particular value even exists, I got many 'undefined' in my string, which can be deleted at the end, after which the *Simplify_Term* function reduces the entire string to complete the entry for the shadow catalog.

We now arrive at function nine, *Search_Catalog*. Obviously, this function requires a string as its input, the search term, which it normally gets from the text box with the ID *searchTip*. The function is triggered by the user, who can click the search button, or hit the Enter key when the focus is on the text box. This last bit is established with the method *addEventListener*, which is a little function that keeps its ears to a specific type of event, and when it hears it go off, it will perform some code. In this case, this code is triggered every time the user releases a key when typing in the text box of the search bar. When the last key was Enter, the listener will engage the *Search_Catalog*-function. Additionally, I provided instructions that if the user has deleted all the text from the text box, the entire catalog should be rendered again. I experimented with letting the *Search_Catalog*-function fire off every time the user pressed a key to get the experience of a live update, so that with every additional letter the user types, the number of entries is restricted. With a catalog of over 1,500 entries, this proved to be too heavy on the calculation side; the code could not be executed fast enough to get a snappy feel. In its current state, there is still not an instant experience of the result, but since the user needs to press the Search button or hit Enter, there is a greater expectancy on the user's side that it can take a fraction of a second for the catalog to update.[8] The search function itself is fairly straightforward. First, it reduces the search term the user provided by means of the *Simplify_Term*-function. Then, it runs through the shadow catalog and sees if the search term is contained in it. If so, that specific entry will be pushed from the catalog (*catData*) into a new object called *searchData*. This way, we can build a subset of the catalog containing only those entries in which the search term is present. Perhaps you wondered why *Render_Table* took as one of its parameters something called *data*. Now you can see that we can give *Render_Table* either the entire catalog (*catData*) or only a subset related to a search term (*searchData*).

The functions ten, eleven, and twelve are fairly similar. I created them by looking up examples of how to arrange objects alphabetically, adapting them to my own situation. For example, I could not be assured that every item had a title, so I built a check for that. Had I not done that, not all pairs would be

[8] Since the search function is embedded in JavaScript and therefore loaded onto the client-side, the experience of searching is still pleasant enough in that no page redirect or refresh is necessary, as would be the case if the search query had to be given to server-side code and then the results given back to the client.

```
  var editorLan = Object.values(texts.editorCatalog)[language];
  var translatorLan = Object.values(texts.translatorCatalog)
[language];
  var noAuthorLan = Object.values(texts.noAuthorCatalog)
[language];

  htmlTableCat +=
    ‘<tr><td><div class=“entry text text-left”><p class=“text
font-weight-bold”><svg onclick=“Popup_More_Info(’ +
    i +
    “, ” +
    language +
    ‘)” xmlns=“http://www.w3.org/2000/svg” viewBox=“0 0 50
50” width=“15px” height=“15px”><path class=“iconModal”
d=“M25,2C12.297,2,2,12.297,2,25s10.297,23,23,23s23-10.297,23-
23S37.703,2,25,2z M25,11c1.657,0,3,1.343,3,3s-1.343,3-3,3
s-3-1.343-3-3S23.343,11,25,11z M29,38h-2h-4h-2v-2h2V23h-2v-
2h2h4v2v13h2V38z”/></svg> ’ +
    data[i].title +
    ‘</p><p class=“text-muted text font-weight-light”>’;
  // Cannot be sure if entry has an author, date, and place,
so must check it first.
  if (data[i].hasOwnProperty(“author”)) {
    htmlTableCat += authorLan + Print_Names(data[i].author);
  } else {
    htmlTableCat += “<i>” + noAuthorLan + “</i>”;
  }
  if (data[i].hasOwnProperty(“editor”)) {
    htmlTableCat += editorLan + Print_Names(data[i].editor);
  }
  if (data[i].hasOwnProperty(“translator”) == true) {
    htmlTableCat += translatorLan + Print_Names(data[i].
translator);
  }
  htmlTableCat +=
    ‘</p></div></td><div class=“entry text text-right”><td><p
class=“text blue”>’;
  if (data[i].issued) {
    htmlTableCat += data[i].issued[“date-parts”];
  }
  htmlTableCat += “<br>”;
  if (data[i][“publisher-place”]) {
    htmlTableCat += data[i][“publisher-place”];
  }
  htmlTableCat += “</p></div></td></tr>”;
}

// 7. Generating more information on item in a pop-up
// Called by (f)Render_Table_Entry
function Popup_More_Info(number, language) {
  // For readability further on, these variables are defined
here.
  var authorLan = Object.values(texts.authorCatalog)[language];
  var editorLan = Object.values(texts.editorCatalog)
[language];
  var translatorLan = Object.values(texts.translatorCatalog)
[language];
  var noAuthorLan = Object.values(texts.noAuthorCatalog)
[language];
  var numberVolumesLan = Object.values(texts.
numberVolumesCatalog)[language];
  var publisherLan = Object.values(texts.publisherCatalog)
[language];
  var placeLan = Object.values(texts.placeCatalog)[language];
  var yearLan = Object.values(texts.yearCatalog)[language];
  var additionalLan = Object.values(texts.additionalCatalog)
[language];
  var callNumberLan = Object.values(texts.callNumberCatalog)
[language];
  var urlLan = Object.values(texts.urlCatalog)[language];
  var closeLan = Object.values(texts.closeCatalog)[language];

  // The div already declared in the HTML is popupInfo, the
div dynamically created is popupModalInfo. All of the dynamic
html is inserted in popupInfo, and then popupModalInfo is
made into a modal (pop-up) wih the package MicroModal.

  htmlPopupInfo =
    “<div class=‘modal micromodal-slide’ id=‘popupModalInfo’
aria-hidden=‘true’>” +
    “<div class=‘modal__overlay’ tabindex=‘-1’ data-
micromodal-close>” +
    “<div class=‘modal__container’ role=‘dialog’ aria-
modal=‘true’ aria-labelledby=‘modal-1-title’>” +
    “<header class=‘modal__header’>” +
    “<h3 class=‘headline’ id=‘modal-1-title’>” +
    searchData[number].title +
    “</h3>” +
    “<button class=‘modal__close’ aria-label=‘Close modal’
data-micromodal-close></button>” +
    “</header>” +
    “<main class=‘modal__content’ id=‘modal-1-content’>” +
    “<p>”;

  if (searchData[number].hasOwnProperty(“shortTitle”)) {
    htmlPopupInfo += “<i>” + searchData[number].shortTitle +
“</i><br>”;
  }
  if (searchData[number].hasOwnProperty(“author”)) {
    htmlPopupInfo +=
      authorLan + Print_Names(searchData[number].author) +
“<br>”;
  }
  if (searchData[number].hasOwnProperty(“editor”)) {
    htmlPopupInfo +=
      editorLan + Print_Names(searchData[number].editor) +
“<br>”;
  }
  if (searchData[number].hasOwnProperty(“translator”)) {
    htmlPopupInfo +=
      translatorLan + Print_Names(searchData[number].
translator) + “<br>”;
  }
  if (
    searchData[number].hasOwnProperty(“author”) == false &&
    searchData[number].hasOwnProperty(“editor”) == false &&
    searchData[number].hasOwnProperty(“translator”) == false
  ) {
    htmlPopupInfo += noAuthorLan + “<br>”;
  }
  if (searchData[number].hasOwnProperty(“number-of-volumes”))
{
    htmlPopupInfo +=
      numberVolumesLan + searchData[number][“number-of-
volumes”] + “<br>”;
  }
  if (searchData[number].hasOwnProperty(“publisher”)) {
    htmlPopupInfo += publisherLan + searchData[number].
publisher + “<br>”;
  }
  if (searchData[number].hasOwnProperty(“publisher-place”)) {
    htmlPopupInfo += placeLan + searchData[number]
[“publisher-place”] + “<br>”;
  }
  if (searchData[number].hasOwnProperty(“issued”)) {
    htmlPopupInfo += yearLan + searchData[number].
issued[“date-parts”][0][0];
  }
  if (searchData[number].hasOwnProperty(“abstract”)) {
    htmlPopupInfo +=
      ‘<hr><p class=“text-success”><b>’ +
      additionalLan +
      “</b><br>” +
      searchData[number].abstract +
      “</p>”;
  }
  if (searchData[number].hasOwnProperty(“language”)) {
    htmlPopupInfo +=
      “<hr>” +
      callNumberLan +
      searchData[number].language +
      “ (” +
      searchData[number].type +
      “)<br>”;
  }
  if (searchData[number].hasOwnProperty(“URL”)) {
    htmlPopupInfo +=
      ‘<a href=“’ +
      searchData[number].URL +
      ‘” target=“_blank”>’ +
      urlLan +
      “</a><br>”;
  }
  htmlPopupInfo +=
    “</p>” +
    “<button class=‘btn-blue’ data-micromodal-close aria-
label=‘Close this dialog window’>” +
    closeLan +
    “</button>” +
    “</main>” +
    “</div>” +
    “</div>” +
    “</div>”;

  document.getElementById(“popupInfo”).innerHTML =
htmlPopupInfo;
  MicroModal.init({
    openTrigger: “data-custom-open”,
    closeTrigger: “data-custom-close”,
    disableScroll: true,
    disableFocus: false,
    awaitCloseAnimation: true,
    debugMode: true
  });
  MicroModal.show(“popupModalInfo”);
}

// 8. Returning all the names for a given category and prints
them in the format ‘first’ - ‘preposition’ - ‘last’.
// Called by (f)Render_Table_Entry and (f)Popup_More_Info
function Print_Names(entry) {
  if (!entry) {
    return;
  }
  // This will be the container for all names. Note that in any
category (author, editor, translator) there could be multiple
persons and each person’s name consist of several elements
  namesString = “”;
  // We are aiming here for a formatting of FIRSTNAME-PREPOSITION-LASTNAME. Further, persons are separated by a comma and the whole list ends with a period.
  for (var name = 0; name < entry.length; name++) {
    if (entry[name][“given”]) {
      namesString += entry[name][“given”];
    }
    if (entry[name].hasOwnProperty(“non-dropping-particle”)) {
      namesString += “ ” + entry[name][“non-dropping-particle”];
    }
    namesString += “ ” + entry[name][“family”];
    if (name == entry.length - 1) {
      namesString += “. ”;
    } else {
      namesString += “, ”;
    }
  }
  return namesString;
}

// 9. Searching keyword and showing results
// Variously called to initiate search (clicking button or hitting Enter) or re-render catalog (if order is changed)
function Search_Catalog(input) {
  // Final result will be a subset of the catalog
  searchData = [];
  // We will not compare search term with catalog, but simplified version of search term with shadow catalog.
  normalizedInput = Simplify_Term(input);
  for (i = 0; i < catData.length; i++) {
    if (shadowCatalog[i].includes(normalizedInput)) {
      searchData.push(catData[i]);
    }
  }

  // Analyze resulting subset; if zero, render entire catalog with heading title ‘no results’
  if (searchData.length == 0) {
    Render_Table(catData, “noResultsTitleCatalog”, lanIndex);
  } else {
    Render_Table(searchData, “foundItemsCatalog”, lanIndex,
searchData.length);
  }
}
document.getElementById(“searchTip”).addEventListener(“keyup”,
function(event) {
  event.preventDefault();
  // 13 is ‘Enter’-key
  if (
    event.keyCode === 13 &&
    document.getElementById(“searchTip”).value != “”
  ) {
    Search_Catalog(document.getElementById(“searchTip”).value);
    // If user deletes search term, render entire catalog again
  } else if (document.getElementById(“searchTip”).value == “”) {
    Render_Table(catData, “againTitleCatalog”, lanIndex);
  }
});
document.getElementById(“searchButton”).onclick = function() {
  if (document.getElementById(“searchTip”).value != “”) {
    Search_Catalog(document.getElementById(“searchTip”).value);
  }
};

// 10. Sort and render search results or entire catalog by
title, putting empty titles at the top
// Called by (e)sortTitleCaption.click
function Indexing_Title_Catalog() {
  catData.sort(function(a, b) {
    if (a.hasOwnProperty(“title”)) {
      firstTitle = a.title;
    } else {
      firstTitle = “”;
    }
    if (b.hasOwnProperty(“title”)) {
      secondTitle = b.title;
    } else {
      secondTitle = “”;
    }
    return firstTitle.toLowerCase().
localeCompare(secondTitle.toLowerCase());
  });
  // New shadow catalog is necessary, to keep actual and
shadow catalog in same order
  Reducing_Catalogue(catData);
}
document
  .getElementById(“sortTitleCaption”)
  .addEventListener(“click”, function() {
    Indexing_Title_Catalog();
    if (document.getElementById(“searchTip”).value != “”) {
      // This ensures only the search results are shown when
ordering
      Search_Catalog(document.getElementById(“searchTip”).
value);
    } else {
      Render_Table(catData, “beginTitleCatalog”, lanIndex);
    }
  });

// 11. Sort and render search results or entire catalog by
author. If no author, then editor or translator. Empty ones
put at top.
// Called by (e)sortAuthorCaption.click
function Indexing_Author_Catalog() {
  catData.sort(function(a, b) {
    if (a.hasOwnProperty(“author”)) {
      firstAuthor = a.author[0][“family”];
    } else if (a.hasOwnProperty(“editor”)) {
      firstAuthor = a.editor[0][“family”];
    } else if (a.hasOwnProperty(“translator”)) {
      firstAuthor = a.translator[0][“family”];
    } else {
      firstAuthor = “”;
    }
    if (b.hasOwnProperty(“author”)) {
      secondAuthor = b.author[0][“family”];
    } else if (b.hasOwnProperty(“editor”)) {
      secondAuthor = b.editor[0][“family”];
    } else if (b.hasOwnProperty(“translator”)) {
      secondAuthor = b.translator[0][“family”];
    } else {
      secondAuthor = “”;
    }
    return firstAuthor.toLowerCase().
localeCompare(secondAuthor.toLowerCase());
  });
  // New shadow catalog is necessary, to keep actual and
shadow catalog in same order
  Reducing_Catalogue(catData);
}
document
  .getElementById(“sortAuthorCaption”)
  .addEventListener(“click”, function() {
    Indexing_Author_Catalog();
    if (document.getElementById(“searchTip”).value != “”) {
      // This ensures only the search results are shown when
ordering
      Search_Catalog(document.getElementById(“searchTip”).
value);
    } else {
      Render_Table(catData, “beginTitleCatalog”, lanIndex);
    }
  });

// 12. Sort and render search results or entire catalog by
year, putting empty years at the top
// Called by (e)sortYearCaption.click
function Indexing_Year_Catalog() {
  catData.sort(function(a, b) {
    if (a.hasOwnProperty(“issued”)) {
      // Invariably, if there is an ‘issued’ key, then the
actual year is stored two levels deeper
      firstYear = a.issued[“date-parts”][0][0];
    } else {
      firstYear = “0”;
    }
    if (b.hasOwnProperty(“issued”)) {
      secondYear = b.issued[“date-parts”][0][0];
    } else {
      secondYear = “0”;
    }
    return parseInt(firstYear) - parseInt(secondYear);
  });
  // New shadow catalog is necessary, to keep actual and
shadow catalog in same order
  Reducing_Catalogue(catData);
}
document
  .getElementById(“sortYearCaption”)
  .addEventListener(“click”, function() {
    Indexing_Year_Catalog();
    if (document.getElementById(“searchTip”).value != “”) {
      // This ensures only the search results are shown when
ordering
      Search_Catalog(document.getElementById(“searchTip”).
value);
    } else {
      Render_Table(catData, “beginTitleCatalog”, lanIndex);
    }
  });
```

compared and sorted, leaving the entries without a title scattered throughout the order of the other entries. Now, with this extra check, all entries without a title are sorted on top. When I discovered this flaw, I not only fixed the code but also decided to return to Zotero and try to give every entry a title, even if there was nothing on the cover. This interplay between coding the catalog and improving the catalog data is to be expected and might occur several times while setting up the catalog. At other times, you will run into issues for which cleaning up or improving the data will not help. For example, the sorting by author name is inherently messy, as multiple persons can be assigned to one item. In such cases, you can, at first, opt to make up the rules that entries should first be sorted by the original author; then, if there is none by the editor, and then, if there is not an editor, a translator.

As a last note, you may be wondering what the code in the first function does. This controls a fancy tool tip functionality when you hover with your mouse over a button or an element. Including this is as easy as including scripts from Popper.js and TippyTip. This, of course, adds some file size to the website when somebody uses it, but in this case, I did not see an issue with that. The only thing that needed manual instruction is when somebody looks at the catalog from a touchscreen device. Since you cannot hover your mouse on a touchscreen device, TippyTip did not respond correctly, so it is better to simply disable the functionality for tablets and smartphones.

Much more functionality could be built. For example, right now, the sorting is very simple. We could imagine that the same button could sort in two directions (A-Z, and Z-A), or that when sorting by year, given an equal year number, the entries are sorted alphabetical by title. Perhaps performance could be optimized to run all of this faster and smoother. This is always project-dependent. Given the relatively small scale of the catalog, I decided that the functionality as it is is enough and adequate.

7 Productivity: Code Editor and Code Repository

Get where you need to be through the path of least resistance. Developing is not our core business. It is just a tool that we use and discard. And what we want is a working, finished product. It can be refreshing to remind ourselves of these maxims. It is all too easy to be blinded by the latest possibilities in web development, which come with a learning curve that will set back your deadline. If it works, it is alright, we do not need to produce perfect code. In case of doubt, rely on more foundational technology, because with them the chances of being supported for a long time and easily fixing your mistakes or imperfections, later on, are high. For most purposes, we do not need to worry about optimization in terms of processing time and download time. If a JavaScript module claims to do something you want, then it probably is best to include and make use of it.

The most important aspect of productivity is making sure you write your code with a good editor. Writing any of these files, HTML, CSS, JSON, or JS could be done in a text editor as simple as Notepad (Windows) or TextEdit (MacOS)—but that would be madness. I wrote my code for the catalog in Visual Studio Code, which is made by Microsoft and free to use. There are other fine applications out there, and in the future, no doubt, will there be new contenders that could make life even more easy. When in doubt, take the seemingly more popular choice. Several reasons make VS Code great. First of all, in VS Code, you do not just open one file, but rather a project, represented by a folder on your computer. All the files in the folder are shown in a list in VS Code, and by clicking on each file, you open that file in a tab. This makes it incredibly easy to code in HTML, CSS, JSON, and JS simultaneously. VS Code displays your code in a color scheme, using one color for each type of thing (a method, attribute, or variable) and setting the background color to something other than white. I prefer a very dark grey as a background color, which is easy on the eyes, especially at night. With one key combination, *Shift+Alt+f*, the program will reformat your code to make it look neat and tidy again, indenting lines of code to show it belongs to a function, *if*-statement or *for*-loop. You can shrink or expand parts of the code; for example, if you write a long function, you can minimize it to its first line only so that you keep an overview of the surrounding code. Speaking of overview, on the right, there is a vertical visualization of the totality of your code, which you can use to browse through it. The real power of a program like VS Code comes in the assistance it gives in writing code. For example, in an HTML-file, only typing the exclamation mark and hitting enter will give you a boilerplate HTML-file, which includes a declaration of the file being HTML, a header with some information pre-filled, and body-tags. Writing a period and a word and then hitting enter will create div-tags with that word as a class, and doing the same with a pound-sign, #, will make a division with the word as its ID, and so forth.[9]

Extensions can be downloaded for the VS Code, which expands its functionality. For example, I use one called Live Server. When you open your index.html and click on Go Live in the bottom bar, your browser will be opened with the website you are creating. Every time you adjust any of the files in the project, the webpage reloads in the browser to give you an automatic update of what the website looks like and how it behaves under the adjusted code. This works even better when you have a second screen to place that browser window in so that on one screen you code, and the other you see the result.[10] Such extensions and other advanced features like debugging make VS Code a really great choice. It is also easy to pick up, but as you progress and become more skilled with coding, you do not need to switch to more professional software: VS Code is a popular choice among many professionals.

More advanced tools to help you have to do with version control. In JavaScript, you can often rely on code that somebody else has already come up with and is willing to share for free. For this, you will need to install packages, just as is the case with Python (see Chapter Seven). What *pip* is for Python, *npm* is for JavaScript. With npm, you can get those packages; just be sure that you have the version you want, and include them in any of your projects.

[9] VS Code does this through integration of Emmet.

[10] I used my iPad as a second screen with the app Duet Display, using a Mounty from TenıDesign to fasten my iPad on the side of my MacBook screen.

The code you write yourself also needs to be controlled, of course, for which currently one of the best options is the software Git, especially as it is offered through the free service GitHub. With GitHub, it should be noted, you do not only get version control but also a reliable way to distribute and collaborate on code. A free account on GitHub will only allow you to make public repositories, so be careful not to use it if you wish to keep your code to yourself. The upside is that with GitHub you also get free hosting and, by adjusting a few settings, your code is not just readable as code in the repository but can be seen working as the website it is supposed to be. This is convenient to demonstrate the project to others and get a reliable sense of how it performs online. You will do well to spend half an hour understanding the philosophy and mechanics behind Git and Github. VS Code supports Git right in the editor, allowing you to see at a glance how you have changed your code since the last save in the repository (which Git calls a 'push').

A note of caution needs to be made about GitHub and, in fact, about all software I refer to: at this moment of writing, they perform really well and are free to use, but both aspects could change over time. GitHub is a particularly good example of this, as its free use and absence of advertisements give off the impression of being part of a public space. The public space on the internet is, in fact, really small. Wikipedia and The Internet Archive are notable examples of stable, future-proof, non-profit organizations committed to keeping their websites open for free, but GitHub is in fact owned by a for-profit company. If it is bought up or pressured by shareholders, it could very well change its services. We have seen such a change of course with Academia.edu. This website purported to create a public sphere for scientists and scholars but shifted in one year, 2016, towards an aggressive strategy to shake money out of their users.[11] With whatever we do, it seems to me to be best to think in terms of abstract ideals and needs and consider which actual technical tool suits you best, keeping an especially close eye to the possibility to exit a certain technology without losing data. In other words: If the company that produces a certain software goes bankrupt, will I able to port my data to other software? Or is it now trapped inside the unstable, unsupported software? You should ideally have a positive answer to this question.

[11] In early 2016, Academia.edu started charging money to have your new publication seen by others as 'recommended', a move so brazen it was mistaken at first for a scam, cf. Ruff, C. "Scholars Criticize Academia.edu Proposal to Charge Authors for Recommendations." *The Chronicle of Higher Education*, January 29, 2016. Later that year, a 'premium' service was introduced, broadly criticized as predatory, cf. Bond, S. "Dear Scholars, Delete Your Account At Academia.Edu." *Forbes*, January 23, 2017.

The last point on productivity has to do with knowledge acquirement. In the end, what will likely be the best way for us, students and scholars, to learn web development is to skim books and apply and fiddle around with actual code. There is, however, a large amount of consumable media such as videos on YouTube and podcasts, which are also helpful. We would not want to spend our working hours on them, but by subscribing to some handpicked channels, you may find yourself using these to fill the time you use to relax: for example, while exercising or while traveling on public transport. The best one to choose, in my experience, aim for a beginners-to-intermediate level and do not aim to cover the latest news but do long-form discussions of best practices. Even if at first you do not understand what they are saying, there is merit in listening. Over time, you will slowly pick up the vocabulary currently in use by web developers, and you will get a sense of what is currently hot and what is not.

8 Quantitative Analysis of the Collection

Having a data set like this in JSON format makes it ripe for quantitative analysis. To get started with that, we can simply look through the list of items in Zotero. We can write some simple calculations ourselves, and we can rely on already developed JavaScript components for more complicated things.

In the current method of cataloging, a total of 1528 objects were registered. The vast majority of them, 93%, are books. The rest are mostly journal articles and, notably, a total of twenty-six manuscripts. To display this properly, we would need a pie chart that first divides 93% books to 7% other materials, with the ‘other materials’ clickable to show a new pie chart dividing up these other materials into 70% journal articles, 24% manuscripts, 4% book sections, and 2% other kinds of documents. Obtaining this information was as simple as sorting the items in Zotero according to kind and then selecting all of one kind to see the number of items.

To extract more specific information, we do not need to go over to Python, but we can insert a few temporary lines of code in the JavaScript we already have. Let us start with looking at the year of publication. At the function *Render_Table*, just before the first for-loop, we can insert *var catchYears*, and inside the loop, we can then write *if* (*data*[*i*].*issued*) {*catchYears* += *data*[*i*].*issued*[*“date-parts”*] + *“v”;*}. Right after the loop, we include *console.log*(*catchYears*). If we run the catalog in a browser, we can then open the console[12] to see the result:

[12] Safari: Develop > Show JavaScript Console. Chrome: View > Developer > JavaScript Console. Firefox: Tools > Web Developer > Web Console.

namely, it will have printed the variable *catchYears,* which is a long string of dates separated by the letter v. This should be copied and pasted into your code editor. With regular expressions, we can break the line at every 'v' (and deleted that 'v,' of course). What is left is a list of more than 1200 lines on each of which is a date. From here, we can use a standard JS library to display bar charts. I used graph.js to make it. I added myself a few lines to give the bars a random color. I also added another package called chartjs-plugin-annotation.js to add a line indicating the halfway point. The result is a bar chart that gives an impression when the books in Geyer's possession were published. Notably, the halfway point lies in between 1902 and 1903, or, in other words, Geyer acquired many books as they were coming hot from the press and invested much less in older books.

Another piece of information is the place of publication. For analyzing this, we need to use uniform names. Thus, Petrograd becomes Saint Petersburg, and Den Haag becomes The Hague. Leningrad was similarly filed under Saint Petersburg to find it alongside the other publications from that city. Christianiae was filed under Copenhagen, and Bulaq and Misr were filed under Cairo. In all those cases, a note in the abstract was included to make the user aware of what is actually on the title page of the item. We could do a similar approach as in the previous paragraph, and this way we could find out how many uniquely different place names are used (96 in the case of Geyer's collection). However, place names also call out for plotting a map to visualize the information. In fact, plotting data on a map is a good example to see how knowing even a little bit of JavaScript can really make a difference. If we would have no skill whatsoever, we would be left with very few options. One possibility would be Palladio, a DH tool developed at Stanford. It wants the following:

- One Excel sheet with one column, indicating the originating city and destination (Vienna), separated by a comma, and 1209 rows.
- One Excel sheet with one column, 96 rows, with unique city name, a comma, longitude and latitude in between quotation marks, and separated by a comma.

Loading this into Palladio gives a pretty looking map with curved lines towards Vienna whose thickness is related to their prevalence in the data set. Besides the idiosyncratic demands on the input, there is no way of getting that map out of the Palladio work environment. They themselves suggest that making a screenshot is the best way. In the spirit of Chapter Four, we see yet again how a well-funded team project has failed to deliver a fully functioning product. With basic knowledge of JavaScript, we can make use of a library called *amCharts*

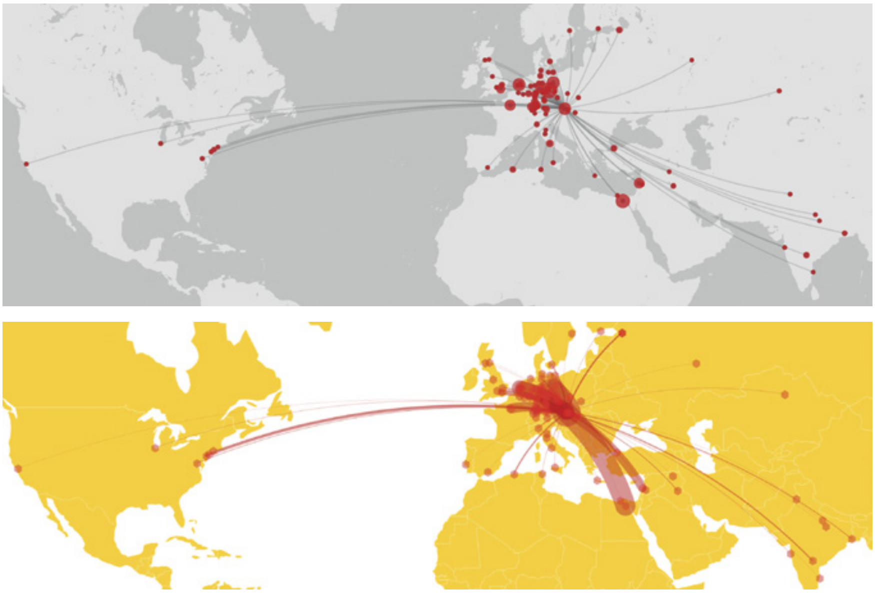


FIGURE 6.1 Above: Map created with Palladio. Below: Interactive map created with amCharts

that gives us much more flexibility and is, in general, just better than Palladio. Of course, amCharts also requires the data to be in a specific format, and the following is a short description of how I changed the place name data as it was in Palladio's requirements towards amCharts preferred format. From the Excel sheet with all the place names, I copied the column of 1209 city names into a new column and selected Data > Table Tools > Remove Duplicates. Then, in a third column, I typed =COUNTIF(A:A;B1). This counts the number of times the value in B1 is to be found in column A. I then selected and copied this cell, pasting it along the C column for as many rows as there were in column B (the list of unique cities). This will get you a unique list of cities together with the number of occurrences. Then, I created the column with latitude and longitude information in column D, mapping it according to the names of column B. For this, I used a website, on which I manually searched for the city name and retrieved its geolocation. I chose this over an automated approach using an API because the number of 96 cities was manageable, and a fair number of them are too obscure to be handled automatically as they would be misidentified or not identified at all. Finally, in column E, I made sure to get the name, location, and occurrence together, like the following:

```
Washington,“38.907192, -77.036871”,4
```

I copied this column into my code editor, Visual Studio Code, which has an advanced Find and Replace function. With the regular expressions function on, with trial and error, I landed on the following Find expression:

```
(([a-z|A-Z]|ü|ö|)+),“(\w.*), (-?\w.*)”,(\w*)
```

And for the Replace expression I took:

```
{\n“id”: “$1”,\n“svgPath”: targetSVG,\n“title”:
“$1”,\n“latitude”: $3,\n“longitude”: $4,\n“scale”:
$5*scaleVar\n},
```

The result can then be inserted into the JavaScript code of amCharts under “images” of the var map. Just above the var map, I made sure to include *var scaleCity = 0.0; var scaleLine = 0.2;*. This will allow us to play around with the city sizes and line sizes according to their occurrence. I have found 0 and 0.2 to be good values for this data set. The result is a dynamic map of the world with dots for each city that produced at least one book that made it into Geyer’s collection, and curved lines towards Vienna whose thickness is commensurate with the number of books coming from that city. This map is zoomable, and when the user hovers the cursor over a city, it will see the city name and the number of items published there.

In the last two chapters, we laid the foundations for working with digitized manuscripts, concluding that we can best store text in a plain text file format such as XML or JSON, and store symbols or shapes in a vector image file format such as SVG, and, lastly, to inform and connect all these matters by academic standards such as TEI and IIIF. In this chapter, we learned of the basic skills to practically put all these files together to create a visual appearance—which we can call a website, web app, digital edition, or digital catalog. The two most important lessons to draw from this and the previous chapters are that technology only remains as powerful as the user wielding it and that we do not need to know everything—just those aspects that help us build a solution. This is why it is very important to keep an open and creative mind. With knowledge of the fundamentals of different technologies, it will be relatively easy to understand how a certain problem can be solved and whether you can learn how to implement that solution in a time that is sufficiently short enough. Now that we have acquired a few basic assets for our practical toolbox, let us, in the next chapter, reach for a higher level of technical skill by delving into the programming language Python and its application to codex images.

CHAPTER 7

Codicology: Automated Analysis Using Python and OpenCV

> Grau, Teurer Freund, ist alle Theorie.
> Und grün des Lebens goldner Baum
>
> GOETHE, *Faust*, 1808. Der Tragödie erster Teil. Studierzimmer, Mephistopheles zum Schüler

Of all aspects of manuscripts studies, codicology has been singled out as relying the most on the actual artifact. Interest in codicology, it is said, will maintain its relevance in having access to the actual manuscript, not just a digital surrogate. So far, advances in 'digital codicology' have confirmed this. I am thinking here of efforts to capture specialized photos that reveal codicological aspects. There are numerous examples of multispectral imaging,[1] hyperspectral imagining,[2] thermographic imagining,[3] and x-ray imaging.[4] Other advanced techniques have also been used, such as scraping off or otherwise

[1] E.g. Arsene, C.T.C., P.E. Pormann, N. Afif, S. Church, and M. Dickinson. "High Performance Software in Multidimensional Reduction Methods for Image Processing with Application to Ancient Manuscripts." pp. 1–25 in *Manuscript Cultures*, 2016; Hollaus, F., M. Gau, R. Sablatnig, W.A. Christens-Barry, and H. Miklas. "Readability Enhancement and Palimpsest Decipherment of Historical Manuscripts." pp. 31–46 in *Kodikologie Und Paläographie Im Digitalen Zeitalter* 3. Norderstedt: Books on Demand, 2015; Easton Jr., R.L., and W. Noël. "The Multispectral Imaging of the Archimedes Palimpsest." pp. 39–49 in *Gazette Du Livre Médiéval* 45 (2004).

[2] Shiel, P., M. Rehbein, and J. Keating. "The Ghost in the Manuscript: Hyperspectral Text Recovery and Segmentation." pp. 159–174 in *Kodikologie Und Paläographie Im Digitalen Zeitalter*. Norderstedt: Books on Demand, 2009.

[3] Meinlschmidt, P., C. Kämmerer, and V. Märgner. "Thermographie—Ein Neuartiges Verfahren Zur Exakten Abnahme, Identifizierung Und Digitalen Archivierung von Wasserzeichen in Mittelalterlichen Und Frühneuzeitlichen Papierhandschriften, -Zeichnungen Und -Drucken." pp. 209–226 in *Kodikologie Und Paläographie Im Digitalen Zeitalter* 2. Norderstedt: Books on Demand, 2010.

[4] Deckers, D., and C. Glaser. "Zum Einsatz von Synchrotronstrahlung Bei Der Wiedergewinnung Gelöschter Texte in Palimpsesten Mittels Röntgenfluoreszenz." pp. 181–90. in *Kodikologie Und Paläographie Im Digitalen Zeitalter* 2. Norderstedt: Books on Demand, 2010.

© CORNELIS VAN LIT, 2020 | DOI:10.1163/9789004400351_009

analyzing little pieces for DNA[5] or comparing ancient paint recipes with the paint analysis of surviving illustrations.[6] All these methods rely on the physical manuscript and do not create a digital surrogate that can be used beyond the single analysis it was meant for.

Perhaps conventual codicology can, after all, benefit from using digitized manuscripts, without needing expensive and advanced technology. In this chapter, we will manipulate digitized manuscripts using the programming language Python and the software library OpenCV (both to be explained below), which are free to use and relatively easy to learn. No team is needed, no grant money is required. Just you, your computer, and a bunch of digitized manuscripts. The plan of the chapter is to walk through an actual project from start to finish. If you actively participate by replicating all of the code and by playing around with it, you will be able to run this code yourself. More importantly, by the end of the chapter, you will be able to customize the code and implement your skills in other tasks both within and beyond codicology. In fact, the most important information in this chapter is not specifically how you can do the case study that I present but to learn of the general principles of programming and the basic strategies to figure out a programmable solution. The explanation of Python and OpenCV is only a case study to learn how programming in general works.

As a case study and proof of concept, we will investigate a unique feature of manuscripts produced in the Islamic world, namely their closing flap (*lisān*). Codices from Europe have a front cover, spine, and back cover. However, Islamic manuscripts very often have an additional flap attached to the back cover that folds over the long edge opposite the spine onto the front, thereby giving the codex more protection and integrity. Some notebooks nowadays have this too, falling on top of the front cover to keep it closed. But classical Islamic manuscripts were designed to tuck the flap underneath the front cover. This flap is called a ‘tongue’ in Arabic, perhaps because of its shape, which is triangular and ends in a tip. With a dataset of several thousands of manuscripts, I set the challenge to automatically detect the angle this triangle makes. This may seem like an uninteresting aspect but that’s exactly the point: who knows what kind of information is contained in such an innocent-looking aspect. Perhaps this

[5] Stinson, T. “Counting Sheep: Potential Applications of DNA Analysis to the Study of Medieval Parchment Production.” pp. 191–207 in *Kodikologie Und Paläographie Im Digitalen Zeitalter 2*. Norderstedt: Books on Demand, 2010; Teasdale, M.D., S. Fiddyment, J. Vnoucek, V. Mattiangeli, C. Speller, A. Binois, M. Carver, et al. “The York Gospels: A 1000-Year Biological Palimpsest.” pp. 1–11 in *Royal Society Open Science* 4 (2017).

[6] Barkeshli, M. “Material Technology and Science in Manuscripts of Persian Mystical Literature.” pp. 187–214 in *Manuscript Cultures* 8 (2015).

angle is always exactly the same, showing some kind of industry-wide agreed upon standard. Maybe it is highly irregular, showing personalized craftsmanship. Such will tell us something about book production in the Islamic world. Maybe there are a couple of standards which can be related to different eras or parts of the world. Such would be valuable information to take into account for manuscripts without a date or place; you measure up the angle of the flap and you have one additional argument to help make an educated guess about the origin.

1 Why Code?

Coding means giving one or more instructions to a computer. Before turning to any programming language as a solution, we need to know if a computer-supported answer is the right solution. There are generally two good reasons. First, there are some things that a computer can figure out which are beyond the capacity of a human being. This can be seen in the case of specialized photography, revealing aspects of a codex that elude the naked eye. Second, there are some things that a computer can figure out for which a human being may not have the patience or stamina. Think of a chess computer that quickly goes through all possible moves and all the moves that each move would possibly result in, and so forth, almost one to two dozen moves ahead: a human being *could* do it, provided they have enough scrap paper and plenty of time. But a computer can do this kind of monotonous work in seconds (grandmasters are typically able to think about five moves ahead). The difference is so huge that the time it takes to write a chess computer that calculates two dozen moves ahead is lesser than to actually calculate two dozen moves ahead yourself. And once it is written, it can be used over and over again. A popular introduction to the programming language Python is called *Automate the Boring Stuff with Python* and this encapsulates quite well the main reason for letting a computer do the heavy lifting for you: it would be too boring to do it yourself. The case study for this chapter falls squarely into this second category. Sure, we could use a *geodreieck* to measure the angle of a flap by hand, write it down, do it again for the next thousand manuscripts, and arrange the resulting angles in a meaningful way, for example, by counting them all up and dividing by thousand to get the average. But it would take an unjustifiable amount of time and its results could be repurposed only in a limited way. Furthermore, quality assurance relies solely on judging the reliability of the person measuring the angle. With a computer running a program, the answer can be obtained much faster, both the program and the results can be used and repurposed in future research, and the quality can be checked by judging the code and running tests and examples with known solutions.

Once you are convinced that writing a computer program is a good technique to get the answer to your research question, you will likely need to adjust your question (and your solution) to make it programmable. At the time that I am writing this, we live in the age of *smart devices*; pieces of ordinary technology beefed up with features that are designed to assist you based on your specific situation. Apple's 'Suggestions' feature and Google's 'Autocomplete' feature literally attempt to guess how you want to finish your sentence. At the core, a computer is anything but smart. It is very dumb indeed, and it will do only and exactly what you tell it to do. Furthermore, it operates on a binary principle: everything must be either yes or no, on or off, one or zero. 'All theory is black or white,' says Goethe, and so rings the paradigm of the age of computers. However, life around us presents itself in all its vivid colors. In other words, we need to convert that colorful real-life problem into a black-and-white digital problem. In our case study, we do so literally. We literally transform the color image of a manuscript cover into a black and white object living only in the virtual memory of a computer, which is deleted the moment the computer finishes running the code. For humanities research, this approach can be frustrating. Instead of saying 'it is' or 'it isn't,' we often times say 'it is kind of,' 'it is similar to,' or 'it could be.' Both our research input and output present the quality of life in all its vivid colors. Take a photo of a manuscript cover, for instance. Classically trained as we are, we instantly recognize the shape of the codex, the surface on which it is resting, the colors, materials, and embellishments. We are so good at this that even if it is severely damaged, we can almost instantaneously make out all these features. We see them as they are and can easily make subjective judgments of the quality, beauty, and even the authenticity of the codex. A computer, on the other hand, sees nothing of this. All the computer 'knows' is the color code for every pixel of which the image is made up. A computer cannot 'see' the manuscript through the pixels, the proverbial forest through the trees.

To figure out if the real-life, colorful question can be turned into a digital, black-and-white question, it is a good practice to collect all information available about our starting position and about the desired end result. A good rule of thumb is to try to transform every bit of qualitative information into quantitative information. Once the start and end positions are clear, you can write down the intermediate steps to figure out how one could get from the start to the end. This does not have to be very detailed, as you will repeat this very same process multiple times on more detailed levels to figure out the actual steps needed between the two steps of the general path. This activity is called 'writing pseudo-code.' Pseudo-code is a rather colorful step-by-step description of what ought to be done, using vocabulary, grammar, and punctuation of a human language such as English. It is the blueprint for writing the actual code, the black-and-white description written in the vocabulary, grammar, and the punctuation of a computer language (such as Python).

2 Description of Case Study

This case study began as a brainstorming session on what to do with access to a couple of thousand digitized manuscripts, all stored as PDFs in which each page contained one photo of a page-spread of the manuscript. Several ideas took root. Chiefly, they were text block analysis, seals analysis, and codex analysis. Text block analysis could consist of measuring its size or its number of lines, which would yield interesting metadata of its own and would be a first step towards automated analysis of the text contained in the manuscripts. Seal analysis would focus on detection of same or similar seals across a corpus. Since seals are mechanically constructed they are supposed to look exactly the same at all times and that seems like an easily exploitable feature for automated searching. Its results will be significant for ascertaining the provenance and ownership of a manuscript. The digital research on them has so far been limited to handmade databases.[7] Codex analysis could focus on medallion-shaped mandorla ornamentation (*lawza, shamsa* or *jāma* in Arabic) or the flap (*lisān*).

A subset of 2500 manuscripts, of the Nuruosmaniye collection, included a photo of the cover as a first page of the PDF. It was decided to focus on flap, specifically its angle. It is possible to extract the angle from a photo without knowing the true dimensions of the codex, since an angle stays the same no matter how you scale the two lines forming it, hence the number of variables is relatively low, making this case study relatively straightforward. Also, it may be noted that this aspect has never been rigorously studied, thereby providing an opportunity to explore a research question never before asked.

At this early stage, a two-step plan emerged: (1) extract the correct image from a PDF file, and (2) analyze the image to find the angle. A logical third step would be to store and process the resulting angle. However, since this was considered to be of lesser difficulty, it was at first left out of the plan.

For step one, I searched for a package (or library) that would handle PDFs within the ecosystem of the programming language Python. A package is a

[7] The example I was thinking of, Chester Beatty's Islamic Seals Database, has been taken offline sometime during 2018. The digital world has precarious life.

pre-written block of code that you can download and use for free. With 'ecosystem' we mean all available packages combined, especially as they are indexed or discussed on major online resources for the particular programming language. For us budding programmers, it is terribly important to leave the heavy lifting to the professionals. PDFs are a prime example of that; once you search around for automatic PDF manipulation, you learn quickly that as nice and shiny PDFs look in a viewer, they can be an absolute mess under the hood. Broken down into its bits and bytes, understanding PDFs is only doable after investing a lot of time, which is why it is important to use a package. There were a few choices, all of them not very well documented, and I landed on PyPDF2, more on which later.

For step two, I then searched for an image manipulation package. I reviewed some candidates, trying to figure out which package was popular, still actively developed, well-documented, and would actually do what I needed. I was convinced by OpenCV, especially after seeing demonstrations of it, for example its use in programming a self-driving car for the popular video game *Grand Theft Auto*. It seemed to have a low barrier to entry while also packing advanced options, assuring me that I would be able to learn it and then keep using it as my use became more demanding.

3 Introduction to Python

The choice for Python for use in digital humanities is easy. From a technical point of view, what Python has got going for itself is that it is a high-level computer language. This means that it looks a lot like actual English, which makes it easy to read and write. It also means that it takes care of a lot of things without requiring the user's explicit command. A notable example is the automatic garbage collection. Your code will likely need to temporarily store virtual files. For example, when we direct our computer to open a PDF file, this PDF is opened in the computer's memory, its RAM. When things are no longer used, Python automatically erases them from the computer's memory, freeing up memory for future tasks. Other programming languages would not do anything until you, the programmer, told them to remove all unused items from memory. Moreover, Python works on pretty much any platform (Windows, macOS, Linux, Android, and so on) which makes it easy to carry over your work, or to integrate somebody else's work.[8] Other programming languages are

[8] The reality is more complicated but also exponentially more technical which is why I do not go into it. This chapter is meant as a primer in Python, not a specific and rigorous introduction.

very close to Python in these regards, a notable example being Ruby. However, a decisive argument in favor of Python is its ecosystem. Every programming language has its own ecosystem of prewritten libraries and packages that you can freely make use of. But because this collection is created by its users, it can have some use cases that are very well covered and others quite poorly. Python has an excellent collection of packages for many of the tasks we would want to perform in the humanities. This is true for our purpose of image manipulation, but perhaps even more so once we have the texts we want to study in digital format and want to perform text analysis on it.

From a practical point of view, we may note the maturity and popularity of Python in general and among scholars of the humanities in particular. This means that Python is a reliable language that will not fall into obscurity anytime soon and render our skills in it useless. It also means that the language itself works wonderfully, without bugs. More importantly, it means that documentation, training, tutorials, and questions and answers are plentiful, in the form of websites, books, and videos. If you do not know how to do something, searching for it online will likely find you with the details of a conversation between somebody who wanted to do (almost) the same and somebody who explains how it can be (or why it cannot be done) on a website like stackoverflow.com. It is worthwhile to become an active member of Python's user community, engaging in current conversations and beginning new ones if your specific question has never been answered. Quite often, you will find that the information you are after is available in different formats, geared to different levels of experience. The popularity further means that there are a lot of packages and example code available on websites like pypi.python.org and github.com. Specific implementations for digital humanities can be found in greater instances than other programming languages, which makes it easier to hit the ground running and share and discuss active projects. Coding means to have a web browser open at all times, on some page of the documentation of Python or OpenCV, on StackOverflow, on some enthusiastic blogger who explains his experiences, or perhaps Wikipedia. Searching, reading, trying, repeat. There is no point in reinventing the wheel.[9]

Let us discuss how to actually *do* Python. Whatever happens, now that you have decided to actually start programming, you should reassure yourself that you have made the right decision. There will be plentiful frustrating moments

[9] At the same time, there will be moments where it will be easier to write your own functionality rather than trying to adapt an existing one. In programming jargon, this is called 'rolling your own'. With experience you will be able to judge the solutions that are currently out there and see if they exactly fit your needs and if not, how much time it would take to make the required changes versus the time it could take to write it from scratch.

when your computer is not doing what you want it to do. For every problem there is a solution, and it is not until a much more advanced stage that you will be the first one to encounter a newfound problem. Since many people have gone before you, they have already found the solution to your problem and talked about it on the internet and in books. Learning a programming language with no real prior experience is not going to happen overnight. So it is good to keep telling yourself that you are not 'losing' time over finding a fix for whatever is going wrong, because this is time invested in better understanding what programming is and how computers can and cannot help you.

To get going requires some initial setting up. Since all of this is a one-off type of work, you might as well have an experienced user (in your local DH discussion group or DH Lab at your faculty or library) do it. Moreover, since the process and requirements may change over time, I shall refrain from a detailed description. What does bear mentioning here is that Python requires *installation*. The way you can have your computer do anything through the Python programming language is by installing some applications from Python.org. They come down to two: a standard library of functions (think: a minimum vocabulary) and an interpreter (think: somebody who listens to your commands in the high-level Python language and tells the computer in a low-level computer language what to do). Once installed, you can write code in the Python language in any plain text editor and save the file as FILENAME.py. There is no graphical user interface for Python. Instead, you use a command line to run the Python interpreter and tell it to read your FILENAME.py. Your computer is now executing your commands!

Confusingly, there are two versions of Python in use. In 2008, Python 3 was released which has been steadily developing. In that same year, a new version of Python 2 came out, 2.6. Version two and three are truly different and the code of one version will not run in the other version. To highlight the popularity of Python 2, even after the release of Python 3.1 (in 2009), Python 2.7 came out (in 2010). That was the final release of Python 2 and is, until today, widely used. I use Python 3 myself and recommend it because it is more future-proof. However, be aware that many examples of code you will find on the internet were written for Python 2. This means that you need to be well aware of which version is installed on your computer and, if both, which version is running (interpreting) your script. Last, be absolutely sure that the packages you install using a package manager like *pip* are installed for the specific version of Python that you are using. You can specify the version of Python like so: *python3 -m pip* followed by what you like to do (most likely *install PackageName,* or *list*, to see which packages are installed). In case of any doubt, seek help from a more experienced user to set things up for you and be sure to document it so that you can use it by yourself later.

If you think writing code in a simple text editor and then running it in the terminal seems awkward, you are not alone. There are tools to make life easier. I recommend using a so-called IDE (Integrated Development Environment) which allows you to write Python code, save it, run it, and even troubleshoot it, and all of this in an application that has a lot more graphical elements than a mere command line. *PyCharm Community Edition* is my current choice, which is a free, easy-to-use, and yet advanced application.[10] PyCharm actually knows the vocabulary, grammar, and punctuation of Python and will help you write your code by suggesting which functions to use and which variables need to be specified, by color coding the grammar of your script, and by indicating where you have made a punctuation mistake. It will allow you to run your scripts with a mere click or keystroke and give helpful feedback in case of an error, enabling you to debug your code more easily. For Orientalists: it supports Arabic flawlessly, which cannot be said of all code editors.[11]

4 Introduction to OpenCV

For our project, you will need to install a couple of packages: PyPDF2, NumPy, and OpenCV. OpenCV is an open source package that has functionality for what is called 'computer vision.' When we look at a photo, we can immediately make out the different objects and their features; it would be hard indeed to see a photo as a mere collection of blobs of color. For a computer, this is the other way around; when a photo is loaded into memory, all that is loaded is a color code for each pixel. The photos I use for this project vary in size, but all of them are built up from more than 3.5 million pixels, sometimes more than 5 million pixels. When a photo like that is loaded in memory, all a computer 'sees' is 5 million color numbers. To be more precise, when OpenCV opens an image, what is loaded into the computer's memory is a matrix with each field containing a vector of three dimensions. That is mathematical speak for what can be described as a table where the number of columns and rows correspond to the number of pixels indicating the image's width and height, as though we drew a fine raster over the image, with each field in that table containing a traffic light. Instead of Red, Yellow, and Green as an actual traffic light, these 'traffic lights' have Blue, Green, and Red. And instead of the three lights following each other, they are all in the same spot and their mixture creates one specific

[10] In this chapter I have not made use of Jupyter but it is also a highly recommended tool to have a more graphically rich interface to write Python, with easy functionality to test and tinker with your code.

[11] For example, the popular code editor *Sublime Text* does not support Arabic correctly.

color. Normally, these blue, green, and red colors are divided into 256 shades (from 0 to 255). As such, sixteen million different colors can be composed from mixing blue, green, and red. In OpenCV, this is called the BGR value. You may have heard of RGB, which is how most software handles color. While OpenCV has it the other way around, there is no point in being upset about this. We simply need to remember this fact when we read out a value of the table that represents the image. For example, if we ask OpenCV the color of a pixel, for example the one that is 100 pixels to the right and 100 pixels down from the top-left corner, we will get an answer that can look like (255,0,0). This means that the pixel is pure blue (and not pure red!).

To manipulate such a matrix by hand would be very tedious and complicated. The power of OpenCV is that there are all kinds of functions that perform sweeping transformations of that matrix. One basic operation that is almost always required before doing other things is to reduce the image to a grey-scale image. Instead of each space in the matrix having a three-color traffic light, the spaces now only have one light, which is defined along 256 shades of grey, 0 being pure black and 255 being pure white. This greatly reduces the complexity of an image and makes available a large number of analyses.

We codicologists are of course not the first to want to automate the analysis of photos. Video surveillance in gas stations, restaurants, streets, and almost anywhere, has given rise to a demand for automatic analysis, for example, to give a live count of how many people are in the store. The automatic processing of forms such as cheques and surveys is another example. OpenCV is the result of many years of development towards that end. Us humanists can piggyback on the advances of this industry. However, you will notice quickly that even though OpenCV has many powerful techniques that require only one or two lines of code, it requires a bit more skill to operate than Python itself. There are only a few tutorials online, many of them are written for the implementation of OpenCV in another programming language (C++), and even the official documentation is sparse and sometimes incomplete. Nonetheless, the power of OpenCV is worth the extra effort.

Instead of opening a digital photo with a viewer to make the photo appear on our screen, we now tell OpenCV to open it and are then able to either perform some sort of detection or transformation. As computers are great at answering yes or no questions, OpenCV is great at answering a black or white question. There is no standard function in OpenCV to detect the angle of a manuscript. But if we can reduce those 5 million pixels to the three pixels that define the corners of the flap, then we can get the angle with some simple mathematics. Again, there is no direct, magic function to get to those three pixels. Instead, we need to gradually reduce the number, and do so in a fashion that takes into account all the irregularities (the vivid colorfulness) of the manuscripts. When we have extracted the image from the PDF, we will therefore first apply a number of transformations aimed at reducing the color image into an image in which the background is black and the manuscript is white, paying particular attention to the flap being as solidly white as possible.

5 Step 1: Extraction of Images

In Python scripts, you can include comments that the computer will skip by beginning the line with a *hash* #. It is a good custom to begin code with a signature and a short remark about the use of the script. This script is meant to do the following: it opens a PDF document, takes a page, and saves the image contained on that page as a JPG image.

A computer reads a script line by line: it knows about the lines it has read but has no idea about the lines ahead. Therefore, if you require packages at some point, it is best practice to have the computer load them in at the very beginning so that you are assured that the computer knows about it and, consequently, can use it. This also makes the script more ordered for yourself. In this case, we will need one function from the IO package which gives additional features for *input/output*, i.e., for handling interactions with files on a hard drive. PyPDF2 is the package we downloaded specifically to interact with PDF files. With the function *PdfFileMerger*, we can create virtual PDF files. Since we do not need more, we only need to load those specific functions.

After the packages, I like to declare any variables that are actually used as constants.[12] A variable is a word that can stand in for some value. Programming languages maintain different types of values, such as a string, integer, or boolean. Python does too; but you do not have to declare the type when you define a variable, it will simply adapt to your use. A wrong type can still cause an error. The simplest example is the *print* command which outputs any selected text. This command cannot take a combination of strings (letters) and integers (numbers). So, if you need to mix them, you first need to convert the integer to string with the *string*() command. Variables are very useful because you can keep updating its value and create a different result because of that new value.

[12] Unlike other programming languages, there is currently no true support for constants in Python. The way to define a constant is to simply define a variable and make sure yourself to never change it.

```
# L.W. Cornelis van Lit, O.P. (c) 2018
# Function for fetching a specific page from PDF and saving
as JPG
# ------------------------------------------------------------

# Dependencies
from io import BytesIO
from PyPDF2 import PdfFileMerger

# A string defining a JPG always starts and ends with these
values
startMark = b”\xff\xd8”
endMark = b”\xff\xd9”

# Predefined path where files are
pathOfFiles = “/Volumes/ExternalHardDrive/FolderName/SubFolderName/”
nameOfCollection = “NURUOSMANIYE”
startNumber = 1000
finishNumber = 2000

# ------------------------------------------------------------
# Check if a file exists and is accessible.
def Accessing_file(filePath):
  try:
    file = open(filePath, “rb”)
    file.close()
    return True
  except:
    return False

# ------------------------------------------------------------

# This function takes one PDF file and extracts one page from
it, to save as JPG.
# Only to be used on PDFs in which each page is only an image.

# Check if file exists. If not, go to next.
def Save_page_PDF(nameOfPDF, pageOfPDF):
  if not Accessing_file(pathOfFiles + nameOfPDF + “.pdf”):
    return
```

We will see examples of this later. Why, then, should we declare variables that never change their value? I do this for readability, and this particular case is a good example. Normally, you would open a PDF document with a program that would display the document on your screen. But this time, we want the computer to open it. This means that it will read the PDF file simply as a long series of characters (a PDF viewer is only translating those characters into differently colored pixels on your screen). Our strategy for obtaining the image from the specific page of the PDF file is to look for a pattern in that series of characters that would indicate that this particular section of the series defines an image. By opening one PDF in a PDF viewer and investigating the elements of a page, I came to the conclusion that the images that made up the PDF were likely JPG images. Searching the internet reveals that JPG images, when read as a long series of characters, always start with *b”\xff\xd8”* and end with *b”\xff\xd9”*. If we can get the specific page of a PDF as a series of characters, we can look for those parts to identify the image. More on the specifics of that strategy later, for now, notice that using *b”\xff\xd9”* in the actual code looks awkward and may be confusing when you revisit the code after many months. Moreover, if you share it with somebody else, it will likely perplex the other person to leave in the seemingly random string of characters *b”\xff\xd8”*. Besides, if we use it more than once, we will be prone to make a typing mistake. Instead, in this code, we can substitute it for a word that makes sense to anybody who speaks English, for instance variable names like *startMark* and *endMark*. All my variables start with a small letter and then each next word is written immediately attached to it but with the first letter capitalized. Such a convention is solely mine and not part of Python’s grammar, but it does help for readability.[13]

After the constants, we declare variables that are variable but only in as much as we will change them by hand, while the code never changes them but only uses them. For this script, the only such variable that needs to be declared is the one defining the path to the PDF file. You may have noticed that we split the path into different components, which will come in handy as we reuse some parts to make it easier to iterate the extraction code over many PDF files and also to construct a meaningful file name for the extracted JPG image. Quotation marks indicate that a variable has a string value; and to use all the variables to put together a path to the PDF file, we need all variables to be a string. You might notice that *startNumber* and *finishNumber* do not have quotation marks and that means they are integer values. We will use them as integers as well as strings and will only convert them to string when needed.

[13] There are style guides that can help you with this. For a deep-dive into writing conventions a good start would be ‘PEP 8’, the style guide for Python.

```
# File exists, so prepare the virtual containers.
  merger = PdfFileMerger()
  virtualpdf = BytesIO()

# Opens PDF and takes out specific page
  with open( pathOfFiles + nameOfPDF + “.pdf”, “rb”) as
sourcePDF:
    merger.append(fileobj=sourcePDF, pages=(pageOfPDF - 1,
pageOfPDF))
    merger.write(virtualpdf)
  merger.close()

# Read the desired page simply in its string of values
  pdf = virtualpdf.getvalue()
  virtualpdf.close()

# Find the start and end of the string defining the JPG
  jpgStart = pdf.find(startMark, 0)
  jpgEnd = pdf.find(endMark, jpgStart)

# Reading out the entire string defining the JPG
  jpgString = pdf[jpgStart:jpgEnd]

# Save the JPG to the hard drive
  with open(nameOfPDF + “.jpg”, “wb”) as jpgFile:
    jpgFile.write(jpgString)

# ------------------------------------------------------------
----

# Loop the function
for i in range(startNumber,finishNumber+1):
  print(“Working on Manuscript” + str(i))
  SavePagePDF(nameOfCollection+str(i),1)
```

pathOfFiles can be left empty or, rather, defined as only two quotation marks.[14] This would make the script look for the PDF file relatively to which folder the Python script is in. In the code shown here, an absolute path to the PDF file is given. The example given here works for macOS. For Windows, the absolute path would start with the drive letter instead of */Volumes*. With some effort, it would be possible to construct variables and code that would be system independent, where all the user has to do is fill in the drive name and folder path and the code would work no matter which platform you are on. However, there is no reason to do that here. In general, you write code until a level of functionality that works for you and no further. Going further would be a waste of our time and likely also of processing power. Since we will eventually run this script over hundreds, possibly thousands, of PDF files, straining the computer becomes a real issue. We are nowhere near a skill level that we can actually and intently optimize our scripts to run the fastest. However, being aware that with every line of code we add processing time already helps. More important, however, is the first point that it would be a waste of our own time to figure out how to add functionality that we in practice will not use. The flip side to this is that we might want to use the script in the future for a different purpose or even share it with others. Two principles can support this; first, to include comments in your code and, second, to set things up flexibly using variables and functions that can easily be altered or swapped in/out. My approach is, then, close to the programming paradigm called *procedural programming*. Other modern paradigms are *object-oriented programming* and *functional programming*. I use the other paradigms as well in an interspersed manner, as we will see. It is a matter of complexity and usability that will drive your exact decision how to code. You will undoubtedly choose the path of least resistance. But do keep in mind that what seems like a quick and dirty fix right now may prove to be a major rewrite later on when you want to reuse or share your code. Indeed, reopening your code after a long hiatus will make it look like someone else's code and you will thank yourself for having used clear and useful variable names and having documented the functionality along the way.

We now come to our actual code, which consists of two functions, one called *Accessing_file* and the other *Save_page_PDF*. For function names, I use the convention of starting with a verb which is capitalized and each subsequent word is separated by an underscore and is written in lowercase. Since

[14] As is often the case, programming allows you multiple paths towards the same goal. For directing Python to the right file you could also try to store the path in a so-called f-string.

'PDF' is an acronym and normally written in capitals, I have done so too for the name of the function.

What are functions? They are basically scripts within a script; several lines of code that are not activated by telling the computer to run the script but which have a name and can be called upon as many times as you like within the script. You can put something in a function and, normally, something (else) comes out. It is quite normal to first simply code line after line to get what you want and at the end tidy everything up by placing it in a function and cleaning up all the variable names. Splitting things into different functions is helpful in distinguishing different tasks. For example, the checking of the soundness of a file is a task different from the extraction of an image from a file. An additional benefit is that within PyCharm CE, you can quickly close/open a function with the +/- sign in the margin to retain an overview of your code. If your code becomes even larger and more complex, you can also split your code over different files and load them in as modules. In Python, you can define a function by writing *def NameOfFunction*(*Variables*):. Take *Accessing_file* as an example. Its function is merely to check if a file exists and is not corrupt. A nonexistent or corrupt PDF file, after all, is useless to us. If we were to try to extract a page from such a PDF, the code would give an error. In that case, what we want to put into the function *Accessing_file* is the path to the file. I wrote the script such that we can use the function both for relative and absolute paths, a decision that we can make later. We can call and activate the function anywhere in our script by writing *Accessing_file*(*PathToAFile*) with *PathToAFile* being an actual path to an actual file, for example *Nuruosmaniye/Nuruosmaniye1.pdf*. In the place where the function is defined, we notice *filePath* placed within brackets; this is a variable, and by calling the function and giving an actual path to a file within brackets, we fill that variable with a value. We can, then, within that function use that variable.

Within *Accessing_file*, I use a bit of object oriented programming. I create an object called *file* and use the *open* function. This itself takes two parameters, a path and a mode. By path is meant the actual path to the desired file, which within our function is represented by the variable *filePath*, so that is what we write. The mode is a more technical aspect: do we want the computer to only have reading or writing access, or both? We only need reading access, so we use *r*. The extra *b* is an even more technical detail that tells the computer to read the file as a binary file, not a text file. Since PDF and JPG files are indeed not text files (more on that in Chapter Five), we want to ensure that the computer interprets those files correctly. Once the object *file* has been created, we can close it again, since all we wanted to know is if the file existed and is able to be opened. It is good practice to explicitly use a *close* command to make sure we are not leaving unnecessary things in memory. Last, if indeed the file is openable, we want to return a value of *True*. If we call the function and give it a path to an existing file that can be opened, we will get an answer of *True*. Had we not wrapped the foregoing lines of code in a *try/except* construction, the code would still give an error if the file or path would not be correct. Therefore, we tell the computer to only *try* to open the file. If that works, it will go on to the next line and close the file, and then move to the next line where it will return to the point where the function was called and return *True*. If, however, it cannot open the file, it will go to the *except* part of the code since an exception has been encountered. There, it will see that it needs to return to the point where the function was called and return the value of *False*. True and false, here, act as a yes or no to the question: is the file accessible? You might still be unsure about how to use this function or if it is even required. We will come to that when we discuss the next function.

The core part of our script is the function *Save_page_PDF*, which needs the name of a PDF file and the page number which we want to extract. In this chapter, we wish to extract the page with the photo of the cover of the manuscript, which is almost always on page one. So we will mostly call this function with 1. But the extra flexibility was easily built and may prove useful later when we have a sudden need to extract all page twos or threes. *Save_page_PDF* begins with an *if*-statement. This is probably the most foundational building block of programming. With an *if*-statement, you can perform certain code if a certain condition is met. You can even add *else-if*-statements (written in Python as *elif*) or a catch-all *else*-statement. In Python, the grammar and punctuation of an *if*-statement is much like defining a function: you begin with a trigger word (for an *if*-statement it is 'if', for a function it is 'def'), then the line must end in a colon. All the lines that should be executed within that *if*-statement (or function) should be indented by one tab relative to the *if*-statement. Thus, if you have functions or *if*-statements nested inside each other, you add additional tabs counting from the start of the line. You do not need to declare the end of a function or *if*-statement; Python knows it from where the code is not indented any longer since it assumes that lines that have the same indentation are in the same block of code. Between *if* and the colon goes the condition that is to be evaluated. A simple condition would be 'if VARIABLEX contains a higher value than a CERTAINNUMBER …'. An *if*-statement is ideal for a yes/no question. In fact, if you simply put a variable in between *if* and :, Python will check if that variable is *True*. You can also invert the same by writing *not* before it. In our code, we put the function *Accessing_file* in between *if* and :. We will thus check if *Accessing_file* returns *True*. In this case, we use the *if*-statement to catch files that are nonexistent or inaccessible. If so, we want to immediately abort and exit the function. So instead of looking for cases where *Accessing_file* gives *True*, we want to look for cases where it gives *False*. This is why we write *if not Accessing_file* … Notice the construction of the path to the file from three components: a static variable, a dynamic variable, and a string.

It is entirely possible to do this differently. For example, right now, *Accessing_file* gives *True* if a file exists and is accessible. But we could just as well make it return *False* and then the *if*-statement would not have needed the *not*. Additionally, there are many other ways to do it too. Your code should work and be understandable. If that is the case, you need not worry too much about which way is better than the other. I think the way I did it here makes it quite understandable. Line 37, 'if not Accessing_file …' is almost actual, understandable English.

Now that we know the file is accessible, we want to extract a page and read out its binary value. After a long process of trial and error, I found out that one way to do it is to use the *PdfFileMerger*() class of the *PyPDF2* package. A class can bring about an object, in this case a virtual PDF, that only temporarily lives in memory. The functions it has are called methods, which make it possible to bring together different pages from different PDFs. Once satisfied with the virtual PDF, you can then save it to your hard drive. One advantage is that these operations can leave the original files intact. The way we will use it is to bring about one such virtual PDF and, subsequently, only open one PDF and only take one page and add it to the virtual PDF. Line 41 creates such a virtual PDF which at that point consists of zero pages. Another container we need is to take that virtual PDF and read it in its raw binary value, for which we use the *BytesIO*() class. There is no way to do this directly from the virtual PDF as the *PdfFileMerger*() class only offers a few options. The closest we can get is to use the *write* method and instead of writing it to disk, we write it to another virtual file, this time made possible by *BytesIO*().

So, we begin by opening the file with the *open* function. We do this within a *with*-statement which has the familiar punctuation of closing with a colon and indenting the relevant lines of code. The *with*-statement is a *read* and *close* function wrapped into one: once the computer has moved away from the *with*-statement it automatically closes the file to free up memory. We use the *append* method to extract our desired page into the virtual PDF. We then use the *write* method to store this virtual PDF in a virtual file which we call *bytePDF*. Because *bytePDF* is made from the class *BytesIO*(), we have the attribute *getvalue*() available, by which we can create a variable called *page* which then simply has the binary value of the page of the PDF.

Now, we still need to make the step from a PDF page to a JPG image. When we open these PDFs, we see every page as one big photo, but we can easily understand that there is some encoding wrapped around that image; every page of the PDF starts by defining it as a page with its dimensions and other characteristics. The photo is only one part of that page. This means that the binary value of the desired page does not only consist of the binary value of the JPG but also of other things related purely to defining the page as a PDF page. We found out earlier that the binary value of a JPG always starts and ends with a certain value. Our strategy to get the photo out of the page is to search for the first occurrence of JPG, by looking where in the binary value of the page the begin-string of a JPG can be found and where the end. The *find*() method requires a string we want to look for and the position from where we want to begin looking for it. Thus, in the binary value of the PDF page, the JPG starts wherever from position zero onwards the string *b”\xff\xd8”* occurs. Using that place to start looking for the end of the JPG, we now have two variables that contain the beginning position and the end position of the JPG within the PDF page. We can then define a variable *jpgString* by taking the entirety of the page of the PDF and slicing off anything before the start of the JPG and everything after the end of the JPG. We end up with the binary value of the JPG of that PDF page. Since we are convinced that this binary value makes up a valid JPG image, we can simply write it to disk as a .jpg. Wrapping that command within a *with*-statement is again to ensure that we close the file after usage. Python allows to ‘open’ a file if it does not exist, provided you use the *w*(rite) mode. In that case, Python creates the file for you.

If we would run the script with only what we have discussed so far, nothing would happen. We need to call the *Save_page_PDF* function, and we need to do so only after we have defined it. Putting all functions together in a useful manner is, therefore, done at the end of the script. For example, we could write at the end *Save_page_PDF*(*“test”,1*) provided that the Python script file is in the same folder as a PDF called *test.pdf*. The result would be the appearance of a *test.jpg* in the same folder, which would turn out to be the photo of the first page. This, however, would only work if *test.pdf* actually consists of one photo per page. We did not include error handling if this is not the case. In other words, this script is built for a specific situation. To make real use of the computer’s power, we can ask it to perform the extraction of a page for hundreds or even thousands of manuscripts. In my case, I had access to 2500 manuscripts spread over a few folders, with each folder containing 500 to a thousand PDFs that all had a predictable name, namely the name of the collection and a number. To make the computer extract the first page of each of them, all that was changing in the path to the file was the number. The easiest way was to use the second-most used tool of programming (after the *if*-statement), namely the *for*-loop. It has a cousin, too, called the *while*-loop. Both execute code over and over again until told to knock it off. As usual in Python, the grammar and punctuation demand the word ‘for’, a condition, and then a colon. While the condition is true, the code that is indented is executed. In our case, we simply let a number run from the value 1000 to 2000 using the *range* function. We do so through the variables *startNumber* and *finishNumber*, so that we can easily adjust things. We need the *+1* as, otherwise, number 2000 would not be executed since the *range* function semantically means “up to” and not “up to and including.” Often times, the variable that takes each of the values in that range is named *i*. Now, we can use that *i* to construct a thousand different commands of the *Save_page_PDF* function that each target a different PDF file. Notice how we turn the *i*, an integer, into a string to create a full filename that the function will turn into a full file path.

I also included a *print* command in the *for*-loop. This is because when you work with very big PDFs of several hundred megabytes, sometimes more than a gigabyte, the code that we just wrote actually needs longer than a mere moment to digest a PDF like that. Since we loop it over many PDFs, the *print* command gives us some sense of where the script is, which gives a hint about how long it will still take to complete. Additionally, if your script gets stuck on a certain file without giving an error code, this *print* command will let you know which file caused it. If you want to exercise even more control, you could include these few lines of code:

```
# This goes at the top
import time
start_time = time.time()

# This goes at the bottom
print(“Finished in %s seconds.” % (time.time() - start_time))
```

This code will give you the number of seconds it took to completely execute the script. The *print* command only gives you a line of text in the console, and after doing something else, it is not saved somewhere. *Print* is used mostly for troubleshooting; if you get an error somewhere, you can start *print*ing values of variables at various stages of your script to see how the script behaves and where it goes off the rails. It is useful to accompany this with an *exit*() command which prematurely ends the script. You can place a *print* and *exit* at the beginning of your script and slowly move through it to see around which point the script turns up an error.

6 Step 2: Analysis of an Image

Running the previous script on my data set resulted in a thousand images of covers of manuscripts. Being new at OpenCV, I slowly built my code from the ground up, with continuous testing. The *print*-command, is excellent for building up code and exploring which approach works and which does not. The equivalent of the *print*-command in OpenCV is the *imshow*-function which will display the image you are working on in a window. Your first exercise, then, should be to simply make an object that loads an image through OpenCV using *img = cv2.imread*(*"PathOfFile"*), then to immediately display that image using *cv2.imshow*(*"NameForWindow", img*) (start your code first with *import cv2*). You will additionally need to write at the end *cv2.waitKey*(*0*), which tells the computer to keep the window with the image open indefinitely (any other number would mean that amount of milliseconds), and *cv2.destroyAllWindows*() which in combination with the *waitKey*-function will close the window with the image if you press any key.

Getting from those few lines to hundreds of lines of code that actually detect the angle of a flap took me months of coding on and off, trying to push my code in certain directions only to notice that it was not working, slowly crafting it into the final state that you see here. This process was driven by two primary impulses; sometimes I looked up more information about the capabilities of packages, such as *OpenCV* and *NumPy*, and stumbled upon extraordinarily powerful features that seemed to be applicable to my project and used them to device steps that got me closer to my goal; sometimes I brainstormed ways to achieve my goal through a number of intermediate steps and then went to look for techniques to make those steps possible. What can make this a long, painful process is that it requires the fuzziness and multiplicity of humanities research but funneled through the precision and unicity of scientific research. Often, when looking at an image of a codex, you can hardly imagine the precise solution, and when staring at your code, you feel boxed in by the capacities of the technology and wish for a more fuzzy and user-friendly approach. For example, OpenCV has a slew of techniques to identify corners, grouped under the term Feature Detection. Since the angle of the flap can (should?) be calculated by finding the three corner points of the tip and the upper and lower points of the flap, it seemed rather grand to be able to identify those corners with one simple line of code. Trying this out instantaneously made clear to me just how black and white theory is and how vividly colorful (photos of) manuscripts are. The 'corners' that OpenCV was 'detecting' were seemingly random points in the image that had virtually nothing to do with the shape of the codex. My first step was, therefore, to reduce the color image to a black

```
# -------------------------------------------------------------

# Digital Codicology: Analyzing Islamic manuscripts.
# This code is specifically to measure the angle the cover
flaps make.
# (c) L.W. Cornelis van Lit, O.P., 2017-2018.

# -------------------------------------------------------------

#Initial requirements.
import cv2
import numpy as np
import os

#User variables
imageStartingNumber = 1091
imageStartingDirection = True
#Either write BW for Black and White processed image to be
displayed, or Color for original to be shown.
imageAppearance = “Color”
maximumEdgePoints = 4
kernelbig = np.ones((10, 10), np.uint8)

# -------------------------------------------------------------

# This function checks if the file exists and if the file is
incorrupt.
# number is manuscript number as given in the filename.
# Direction is for browsing when encountering a faulty image.
# True means browsing up, False means browsing down.
# Corruption is checked by making sure file size exceeds 100kb.
def Check_image_readable(imageNumber, direction):
  if direction:
    imageNumber = imageNumber + 1
  else:
    imageNumber = imageNumber - 1
  fileName = ‘NURUOSMANIYE/NURUOSMANIYE{0}.jpg’.
format(imageNumber)
  if os.path.isfile(fileName):
    fileSize = os.path.getsize(fileName) >> 10
```

and white image in which the flap became clearly visible and distinct from the supporting surface. From there I needed to slowly reduce the number of possible points to one: the tip of the flap, all the while leaving room for exceptions, anomalies, and irregularities.

In this code, we will need three packages, and so we import them first. Next, we set up some variables that determine the initial image that should be analyzed. When cleaning up my code, I split it into four functions, each covering a specific part of what the computer needed to do. We first check if the file is an actual and complete image, then we analyze the image to get to the flap, then analyze the flap to get to the angle, and last we want to display our results so that we can get visual confirmation.

6.1 *The Function* Check_image_readable

This function makes use of two observations. One is that all our files have the same name except for a number, the other is that files under about a 100kb are corrupt. We do not want to perform an analysis on a corrupt image. Instead, we would like the computer to move over to the next image. If we would run the angle detection code on only one image per turn, we could have chosen to simply give back an alert when an image is corrupt instead of moving over to the next image. In this case, we wish to see how the script performs under multiple images since there is a lot of variation in the photos and manuscripts, and running it on only one image would possibly keep us blind to edge cases that our code does not facilitate. We use a variable *direction* to indicate which way the computer should cycle; to a number higher or a number lower. Here, I have chosen to do so by making *direction* either True or False. The *if*-statement speaks for itself: 'if direction is true', then evaluate the manuscript whose filename is one number higher; 'if direction is false', then evaluate the manuscript with one number lower. Next, we establish the path to the file by the variable *fileName*. As you may notice, I have hard-coded the path to the file, since this script is in first instance meant to test our setup. To productively make use of this script, we would do well to make this more dynamic by defining *fileName* from the combination of several variables defined at the beginning, as we did in the previous script. In a next *if*-statement, we first check if there even is a file like we just defined, using the *os* package. I chose this function over the *open*-function since it gives us an obvious *True/False* dichotomy and since we will use the *os* package anyway in order to gauge the file size of that file using *getsize*. This *getsize* command gives an answer in bytes, which we can change to kilobytes by writing *>> 10*. Then, we evaluate that file size: if it is over 100, we let the computer go on to the next function; if it is not, we let the computer start over from the beginning of *Check_image_readable*. This would not

```
    if fileSize > 100:
      Analyze_image(imageNumber)
    else:
      Check_image_readable(imageNumber,direction)
  else:
    Check_image_readable(imageNumber,direction)

# -----------------------------------------------------------

#This function analyses the image.
def Analyze_image(imageNumber):
  img = cv2.imread(‘NURUOSMANIYE/NURUOSMANIYE{0}.jpg’.
format(imageNumber))
  gray = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
  small = cv2.resize(img, (0, 0), fx=0.8, fy=0.8)
  smaller = cv2.resize(small, (0, 0), fx=0.5, fy=0.5)
  originalImage = smaller
  height, width, _ = originalImage.shape

  ret, thresh1 = cv2.threshold(gray, 180, 255, cv2.
THRESH_BINARY_INV)
  scaled = cv2.resize(thresh1, (0, 0), fx=0.8, fy=0.8)
  openimg = cv2.morphologyEx(scaled, cv2.MORPH_OPEN, kernelbig)
  ret, thresh2 = cv2.threshold(openimg, 1, 255, cv2.
THRESH_BINARY)
  scaledagain = cv2.resize(thresh2, (0, 0), fx=0.5, fy=0.5)

  _, contours, _ = cv2.findContours(scaledagain, cv2.RETR_
EXTERNAL, cv2.CHAIN_APPROX_SIMPLE)
  cv2.drawContours(originalImage, contours, -1, (0, 255, 0), 1)
  mainContour = max(contours, key=len)

  hull = cv2.convexHull(mainContour, returnPoints=False)

  hullContourX = []
  hullContourY = []
  for indexHull in hull:
    hullContourX.append(mainContour[indexHull[0]][0][0])
    hullContourY.append(mainContour[indexHull[0]][0][1])
```

result in analyzing the same image again, since we change the *imageNumber* at the beginning. Thus, if the computer hits a corrupt image, it will cycle upward or downward, depending on the direction, long enough to find a good image. Notice that the script does not handle the case in which the computer would cycle out of the range of manuscripts. For example, if in the folder there are files with MANUSCRIPTNAME500 to MANUSCRIPTNAME1000, if the computer goes to 499 or 1001, it would theoretically search for a valid image indefinitely. In practice, it will produce an error message stating that the maximum recursion depth has been exceeded. We did not include guard rails for these cases as this script relies on the user to not enter those error states but instead stay within the boundaries of the files.

6.2 *The Function* Analyze_image

With the knowledge that the image is valid, we now begin by reading in the image. Many of the variable names in this function are short and not terribly descriptive, and that is because their use is limited to a specific place in the script; if you understand the context, you understand the meaning of the variable. In this function, the first step is to turn the image into a greyscale image. We also wish to preserve a copy of the original for display purposes. Since the images I worked with are way too big to display fully on a screen, the script includes commands to resize them. Since the resizing of the greyscale image is done in two stages and may have an impact on the process of making the manuscript white and the background black, I resize the original as well in two steps. This may be a part of the script that is ripe for improvement, but in here lies a lesson for us: although a clean and lean script enhances its usability, there comes a point after which cleaning up does not give us that much more usability that it merits the time investment. In this function, after the resizing, we then want to know how big the image has become. Of the different ways there are to get to this information, we use here *NumPy*'s *shape* command which returns the height, width, and number of channels. To store each of them separately, we define three variables at the same time, separated by a comma. Since we only require the first two, we can forget about the last one by inserting an underscore.

The next few lines of code transform the image into a black background and a white codex. This is to some extent an art, not a science. This means that to some extent, you are simply left to play around with some of the available tools and their settings until you get the desired result. Furthermore, this also depends on the images you have. With the test data used here, a particular difficulty presented itself in that the surface on which the manuscripts were lying had a prominent pattern and was relatively close in color to some of the

```
#Hand over output to function that displays image
  Display_image(imageNumber, originalImage, scaledagain,
hullContourX, hullContourY)

# -----------------------------------------------------------

def Find_angle(hullContourX, hullContourY, imageWidth,
imageHeight):
  # Using NumPy requires arrays
  reducedXArray = np.array(hullContourX)
  reducedYArray = np.array(hullContourY)

  # Using NumPy to get index numbers of all Y values within
15% of middle
  reducedYLength = len(hullContourY)
  varianceY = imageHeight * 0.15
  minimumY = imageHeight / 2 - varianceY
  maximumY = imageHeight / 2 + varianceY
  decideMiddlePoints = np.zeros(reducedYLength, dtype=np.int)

  for correctYvalue in range(0, reducedYLength):
    if minimumY <= hullContourY[correctYvalue] <= maximumY:
      np.put(decideMiddlePoints, correctYvalue, 1)
      xpos = hullContourX[correctYvalue]
      ypos = hullContourY[correctYvalue]

      # Ruling out spurious points on non-flap side by
checking if nearby points in terms of x-value are near edge
in terms of y-value (i.e. in a corner)
      # Also ruling out flaps with distortions, i.e. deleting
all middle points on the left/right side of the image where
there is a point which meets above requirement.
      for otherXvalues in range(0, reducedYLength):
        if (xpos - 5 <= hullContourX[otherXvalues] <= xpos
+ 5) and (hullContourY[otherXvalues] < imageHeight * 0.2 or
hullContourY[otherXvalues] > imageHeight * 0.8):
          if hullContourX[otherXvalues] <= imageWidth/2:
            np.put(decideMiddlePoints,np.where(reducedXArray
<imageWidth/2),0)
```

codices. With a clean black background, things will obviously be much easier. *Threshold* is a useful function which looks for all pixels with a value higher than a certain limit and gives all these pixels the same value. To remove noise and make the silhouette of the codex smoother, we can use a morphological function called *opening*. This made it necessary to threshold again, but this time to catch every non-pure black part and make it pure white. For a reason that I did not look deeper into, after a threshold, I first needed to scale the object even if only to the same scale in order for other functions to operate on the object. This is an example of finding a solution to a problem that seems to run way too deep for our level of expertise: if a certain band-aid works to counter the symptoms of a problem, then it might be alright to actually use that band-aid, as the actual cause of the problem might be too complicated to solve or even to understand. Even if the root cause was fairly easy to figure out and correct, there is not much lost here to use a *scale* function merely to get things to work. Of course, there is a small-time penalty when executing the code, but we 'only' work with thousands of files and do not need a super-powered mainframe to crunch our numbers. Even if our own computer will have to crunch an extra second per manuscript (it won't), that should be perfectly acceptable to us.

Now that we have the manuscript as a white object, we can ask OpenCV to find the contours of all white objects. If you experiment with the scripts as you are reading this, you will notice that some manuscripts will not be one solid white blob but have black spots and sometimes isolated small white spots within the black spots. In other words, some parts of the manuscript have been effaced and we will have to take this into account as we move ahead. You may have also been puzzled as to why we created the objects white and the background black since our human eyes would recognize objects much better if the object(s) are black on a white surface. This is not so for OpenCV. The *findContours*-function, which we use, finds white objects on a black background. On the next line of code, we command the computer to draw the contours in bright green. I made a promise earlier to keep functionally different parts separated which would mean that, in this function, we only analyze the image while later taking care of displaying the image. You could argue, therefore, that this is an imperfection in the code. I kept it in to show you my workflow: while developing this, I followed the *drawContours*-function with an *imshow*-command to display the image and check to see if the code generated a desirable outcome.

Since there can be multiple contours, I assume that the largest contour must be the contour of the actual codex. This assumption will produce a couple of so-called type II errors, i.e., false negatives: for images where the flap has become its own white blob, the computer will fail to find a flap (even though it

```
          else:
            np.put(decideMiddlePoints, np.where(reducedXArray
> imageWidth / 2), 0)
          break

  # Only proceed if a flap can be found.
  # Get X,Y value of tip-point, making sure there are some
points that could be the tip

  if not np.count_nonzero(decideMiddlePoints) == 0:
    # Get X,Y value of only tip-points
    tipOfFlapYArray = reducedYArray[decideMiddlePoints != 0]
    tipOfFlapXArray = reducedXArray[decideMiddlePoints != 0]
    restOfHullXArray = reducedXArray[decideMiddlePoints == 0]
    restOfHullYArray = reducedYArray[decideMiddlePoints == 0]

    # tipX needs to be decided, left or right
    # Is the tip on the left?
    if np.average(tipOfFlapXArray) < imageWidth / 2:
      # Get X,Y value of furthest tip-point
      reducedXMinimumArray = np.where(tipOfFlapXArray ==
tipOfFlapXArray.min())
      reducedXMinimumList = reducedXMinimumArray[0].tolist()
      tipX = int(tipOfFlapXArray.min())

      # tipY needs to be estimated as an average
      tipY = 0
      for tipCoordinate in reducedXMinimumList:
        tipY = tipY + tipOfFlapYArray[tipCoordinate]
      tipY = int(tipY / len(reducedXMinimumList))
      # Draw the tip-point
      # cv2.circle(imageForDisplay, (tipX, tipY), 6, (0, 50,
250), -1)
    # Is the tip on the right?
    elif np.average(tipOfFlapXArray) > imageWidth / 2:
      reducedXMinimumArray = np.where(tipOfFlapXArray ==
tipOfFlapXArray.max())
      reducedXMinimumList = reducedXMinimumArray[0].tolist()
      tipX = int(tipOfFlapXArray.max())
```

is there). Since we are working with large numbers, false negatives are always more desirable than false positives. Err on the side of caution.

With the main contour of the codex found and defined as five hundred to a thousand points, our hunt for the three points defining the tip and both edges of the flap has drawn closer. In one swoop, we can reduce this to about twenty to forty by letting OpenCV draw a convex hull around the contour. A convex hull is the shape that encapsulates the entire codex in the least possible points. It is not the surface area that is minimized, but the number of corners. So, within the hull, there may be some 'empty space' between the border of the hull and the border of the contour. In the case of Nuruosmaniye 1091, the entire photo consists of just more than 4 million points, the contour of the codex consists of 782 points, whereas the hull consists of 23 points. It is now our job to find the three points of those 23 that define the flap, or to begin with just the one that defines the tip.

Before we go into that, we let the *Analyze_image* function do one more thing and that is to store the X and Y coordinates of the hull points in lists. A list is much like a vector or an array: one object that contains several values in an ordered manner; much like a bus when looked from the side, behind every window sits somebody (or nobody). They all sit in the same bus, all have a specific place, and yet each is different. When we speak of variables, we usually mean a word used as a name that stores only one value (a string, an integer, or boolean), such as *name = "Cornelis"*, or *publicationYear = 2020*, or *published = True*. With lists, we start to group together a bunch of such values into one object. The definition is the same as single variables; we simply type a word (the name of the object) followed by an equal sign followed by the data we want that word to hold, such as: *divisionOfBook* = [*"Introduction", "Theory", "Practice", "Conclusion", "Postscript"*]. Everything between the square brackets now belongs to *divisionOfBook*. A list is only one type of such collection of values, and the general term for such types of objects is 'data structure.' This technical term can be helpful to understand what is going on: the data is a number of values, and the structure is the way it is stored. For a list, then, the structure is simple: all values are stored one after the other, and the order in which they are stored is preserved. This order-preservation is very useful, because it means that in Python, the value of a specific place can be obtained by calling the name of the object and using an index number with square brackets, such as *divisionOfBook*[*1*], which will return "Theory" (note that Python starts counting at zero.)

When we used the *convexHull*-function we actually got back the index of each point of the contour that together makes up the convex hull. So, with that index and the contour, we can extract the X-values and Y-values. First, we

```
      # tipY needs to be estimated as an average
      tipY = 0
      for tipCoordinate in reducedXMinimumList:
        tipY = tipY + tipOfFlapYArray[tipCoordinate]
      tipY = int(tipY / len(reducedXMinimumList))
      # Draw the tip-point
      # cv2.circle(imageForDisplay, (tipX, tipY), 6, (0, 50,
250), -1)
    else:
      #print("Middle points undeterminedly left or right.")
# It is a bit hacky but we need to specify all seven return
values as None. If we decide later to add more return values,
more 'None' values need to be added here
      return (None, None, None, None, None, None, None)
  else:
    #print("No middle points found!")
    return (None, None, None, None, None, None, None)

  # Check if hull is formed correctly
  if (abs(restOfHullXArray - tipX) < 4).any():
    print("Some point of hull too close to X value of tip.
Computer likely did not identify hull correctly.")
    return (None, tipX, tipY, None, None, None, None)

  # Now find points along the edge and calculate angle

  tipUpperX = tipX
  tipUpperY = np.min(tipOfFlapYArray[reducedXMinimumList])
  tipLowerX = tipX
  tipLowerY = np.max(tipOfFlapYArray[reducedXMinimumList])

  #In the case of e.g. NURU1575 the middle-point-finder could
only find one point, so we need to search by hand for the
other tip-point.
  if len(decideMiddlePoints[decideMiddlePoints==1]) == 1:
    decideOtherTip = np.where(imageWidth - reducedXArray < 4)
[0].tolist()
    if decideOtherTip:
      tipUpperY = np.min(reducedYArray[decideOtherTip])
```

simply make new, empty lists called *hullContourX* and *hullContourY*. Then, we loop through the indices that the *convexHull*-function created and take from the contour of the codex the X-value and the Y value. The *hull* is wrapped twice in a list, meaning that each index is not stored as an integer but as a list containing one element which is an integer. So to get to the correct position in the list of *mainContour*, we should not call *mainContour*[*ElementInHull*], but *mainContour*[*ElementInHull*[*0*]] instead. In Python, counting starts at zero; so, the first element in a list can be reached by [0]. For *mainCountour*, the same is true; each point is not just two values representing the X and Y coordinates, but each point has a list in which there is one element, namely a list of two values. In this manner, we need to add [0] just to get to the X,Y-value pair. Since X is first in the list and Y second, we can get to X by again adding [0], and we get to Y by adding [1]. We use the *append*-method to add those values to the *hullContourX* and *hullContourY* lists. To understand these last couple of technical moves, it might be worthwhile to read the code and the explanation again, carefully, and to play around with it yourself by placing a few *print* commands and an *exit* command right after and then running the script. With that, this function has played its part and needs to hand over its findings to the next function. Of course, we should reasonably expect this to be the *Find_angle* function. However, in the course of developing this code, I came to enjoy it better to immediately parse it to *Display_image* and let that function call upon *Find_angle*.

6.3 *The Function* Display_image

This function is all about visually presenting the information. In the end, we will be able to display the original image, a contour of the codex, the points that define the hull, the tip and the corners of the flap, and a message telling us the angle. I have added a rudimentary user interface in which pressing the key E makes one switch the view from the original manuscript photo to the white-on-black image as processed by OpenCV. Additionally, pressing the key Q makes one switch to the photo one number down, and pressing the key W will switch to the photo one number up.

The first task is perhaps counter-intuitive: to make a color image out of a greyscale image. This does not mean that the photo itself turns colorful again, it remains a greyscale image (or rather, white on black now that it has undergone certain transformations). But it simply means we tell OpenCV we want three channels per pixel again instead of one. This is needed so that we can draw on the image in color, such as the points that define the hull. Were we to draw those points on a greyscale image, they would be grey too and they

```
      tipLowerY = np.max(reducedYArray[decideOtherTip])
      restOfHullXArray = np.delete(restOfHullXArray,decideOth
erTip)
      restOfHullYArray = np.delete(restOfHullYArray,decideOth
erTip)

  #For left flap we need to look at minimal X values, for
right flap maximal X values.
  if tipX < imageWidth / 2:
    # Get all XY that are in the upper quadrant
    upperRestOfHullXArray = restOfHullXArray[restOfHullYArray
< imageHeight / 2]
    upperRestOfHullYArray = restOfHullYArray[restOfHullYArray
< imageHeight / 2]

    upperRestOfHullYArray = upperRestOfHullYArray[upperRestOf
HullXArray < imageWidth * 0.2]
    upperRestOfHullXArray = upperRestOfHullXArray[upperRestOf
HullXArray < imageWidth * 0.2]

    # Get all XY that are in the lower quadrant
    lowerRestOfHullXArray = restOfHullXArray[restOfHullYArray >
imageHeight / 2]
    lowerRestOfHullYArray = restOfHullYArray[restOfHullYArray >
imageHeight / 2]

    lowerRestOfHullYArray = lowerRestOfHullYArray[lowerRestOf
HullXArray < imageWidth * 0.2]
    lowerRestOfHullXArray = lowerRestOfHullXArray[lowerRestOf
HullXArray < imageWidth * 0.2]

    def GetUpperPointNumber():
      return((np.where(upperRestOfHullXArray ==
upperRestOfHullXArray.min())[0]).tolist())

    def GetLowerPointNumber():
      return((np.where(lowerRestOfHullXArray ==
lowerRestOfHullXArray.min())[0]).tolist())
  else:
    # Get all XY that are in the upper quadrant
```

would not visually stand out. We also calculate the height and width. We had done so before, but variables defined within a function only live within that function and cannot be accessed outside of it. We could have parsed the values into this *Display_image* function, but recalculating them seemed just as convenient.

In the next block of code, we encounter one way that a variable is able to be accessed in places other than where it is locally defined, and that is to put *global* in front of it. I did this so that *imageAppearance* would be recognized wherever. The *if-else*-statement speaks for itself: it looks at the variable defined at the very beginning and either makes *imageAppearance* the image processed by OpenCV (white codex on black background) or the image as it originally can be found on your hard drive. If the variable is set to anything else, the code will not execute, and a custom error message is printed to the console.

In the next block, we use a *for*-loop to make all the corners of the hull visible. A *for*-loop is convenient to write these commands in an easy, compact form and also has the added benefit that it is dynamic: whether the hull has 20 points for one image or 24 points for another image, it will iterate over all the points. I figured the corners can be best made visible by points. So, we will draw very small circles which we can do with the aptly termed *circle*-function of OpenCV. This functions needs to know the image it should draw on, the X,Y coordinates (counting from the top-left corner) of the center of the circle, the radius of the circle, the color, and the thickness of the outline. Using –1 as the thickness tells OpenCV that the circle should not have an outline but should have a solid fill. For every point of the hull, we repurpose the variable *xPos* and *yPos* to set the X and Y position of that point, and then draw a circle on it. I made an index on the fly, starting from zero and ending in whatever the length of either *hullContourX* or *hullContourY* is (they are obviously the same length). I am sure there are other ways to go about drawing point-like circles for each corner of the hull, but I found this approach simple (only four lines of code) and understandable.

Next, the function *Find_angle* is called. This function spits out several answers: the angle, the X and Y coordinates of the tip of the flap, and the X and Y coordinates of points along the upper edge of the flap and the lower edge of the flap, which are most often the actual corners of the flap. All these answers need to be filled as a value of a variable within the function of *Display_image*. We can simply do so by defining these variables as the function, separated by commas, of course, according to the order of the answers that the function gives back. For the function to do its magic, it needs four pieces of information: the corner points of the hull in its X and Y coordinates, the width of the

```
    upperRestOfHullXArray = restOfHullXArray[restOfHullYArray
< imageHeight / 2]
    upperRestOfHullYArray = restOfHullYArray[restOfHullYArray
< imageHeight / 2]

    upperRestOfHullYArray = upperRestOfHullYArray[upperRestOf
HullXArray > imageWidth * 0.8]
    upperRestOfHullXArray = upperRestOfHullXArray[upperRestOf
HullXArray > imageWidth * 0.8]

    # Get all XY that are in the lower quadrant
    lowerRestOfHullXArray = restOfHullXArray[restOfHullYArray
> imageHeight / 2]
    lowerRestOfHullYArray = restOfHullYArray[restOfHullYArray
> imageHeight / 2]

    lowerRestOfHullYArray = lowerRestOfHullYArray[lowerRestOf
HullXArray > imageWidth * 0.8]
    lowerRestOfHullXArray = lowerRestOfHullXArray[lowerRestOf
HullXArray > imageWidth * 0.8]

    def GetUpperPointNumber():
      return((np.where(upperRestOfHullXArray ==
upperRestOfHullXArray.max())[0]).tolist())
    def GetLowerPointNumber():
      return((np.where(lowerRestOfHullXArray ==
lowerRestOfHullXArray.max())[0]).tolist())

  # Establishing upper edge points
  upperEdgePoint = []
  upperEdgeX = []
  upperEdgeY = []
  for pointNumber in range(0, maximumEdgePoints):
    if pointNumber < len(upperRestOfHullXArray):
      upperEdgePoint.append( GetUpperPointNumber()  )
      upperEdgeX.append( upperRestOfHullXArray[upperEdgePoint
[pointNumber][0]] )
      upperEdgeY.append(upperRestOfHullYArray[upperEdgePoint
[pointNumber][0]])
```

image and the height of the image (that is, not the original image on file but the image as it is processed by OpenCV). We will discuss the *Find_angle* function later.

Then, the first thing to check is if an angle was found. If no angle was found, the message we want to display is that no flap was detected. If an angle was detected, we want to set the variable *message* to that. I used a feature of Python to include a variable in a string, by placing curly brackets around an index number (starting with zero) and by following the string with *.format*(*VariableName*). If there is an angle, we also want to make the tip and the edges of the flap visible; so we draw points, this time bigger and in other colors than when we drew points for the corners of the hull.

Next, we want to display the *message*. We do this outside of the *if*-statement because we do this no matter the outcome. If you observe closely, you will see that I have OpenCV draw the same message thrice. The first two times, the message is black and the third time, it is purple-pink. I do this so that the purple-pink text looks like it has a black border around it which enhances the readability.

All that is left to do is give OpenCV the command to display the image on our monitor, followed by a command to freeze it until a keystroke. Any key other than Q, W, or E will close the window. By evaluating the keystroke using *ord*, we can make the code readable for humans, as it now simply reads 'if [the] key [stroke] is q …'. When we press Q on our keyboard, we want the computer to analyze the manuscript whose number is one less than the current one. The easiest way is to close the current window and let the script run all over again but this time with *direction* set to *False*, that is, 'downward'. For W, we do the same but with *direction* set to *True* or 'upward.' For pressing E, we want to switch between normal view and analyzed view, which we can do by only letting the last function, *Display_image*, run again but with the variable *imageAppearance* changed. At the moment, we only have two different views, normal and analyzed. So we could have made *imageAppearance* a boolean variable, switching between *True* and *False*. For readability and for allowing future expansion into other views, I thought it wiser to have *imageAppearance* be a string which should be either 'BW' or 'Color.'

The script finally ends with a call to the function *Check_image_readable*, to set everything in motion. I placed a minus one after *imageStartingNumber* since *direction* is at first set to *True* and will, therefore, increase the *imageNumber* by one in the *Check_image_readable* function. All that is left to discuss for this script is the function that figures out the angle based on the corner points of the hull.

```
      upperRestOfHullXArray = np.delete(upperRestOfHullXArray,
upperEdgePoint[pointNumber])
      upperRestOfHullYArray = np.delete(upperRestOfHullYArray,
upperEdgePoint[pointNumber])

  # Establishing lower edge points
  lowerEdgePoint = []
  lowerEdgeX = []
  lowerEdgeY = []
  for pointNumber in range(0, maximumEdgePoints):
    if pointNumber < len(lowerRestOfHullXArray):
      lowerEdgePoint.append( GetLowerPointNumber() )
      lowerEdgeX.append(lowerRestOfHullXArray[lowerEdgePoint[
pointNumber][0]])
      lowerEdgeY.append(lowerRestOfHullYArray[lowerEdgePoint[
pointNumber][0]])
      lowerRestOfHullXArray = np.delete(lowerRestOfHullXArray,
lowerEdgePoint[pointNumber])
      lowerRestOfHullYArray = np.delete(lowerRestOfHullYArray,
lowerEdgePoint[pointNumber])

  # !!! Note that upper/lowerRestOfHull are now empty !!!

  # Calculate degree of angle based on lower edge points
  # First apply Theorem of Pythogaras to find all lengths,
with A = horizontal, B = vertical, C = diagonal

  upperAngles = []

  for pointNumber in range(0,len(upperEdgeX)):
    # First apply Theorem of Pythogaras to find all lengths,
with A = horizontal, B = vertical, C = diagonal
    A = abs(tipUpperX - upperEdgeX[pointNumber])
    B = abs(tipUpperY - upperEdgeY[pointNumber])
    C = A**2 + B**2
    # Then use Law of Cosine to find angle. Answer is
returned
in Radians so it needs to be transposed to Degrees for human
readability.
```

6.4 *The Function* **Find_angle**

We have reduced our humanities problem of finding out more about the flap to a mathematical problem of finding a triangle based on a couple of dozen coordinates. What we need to do, then, is figure out regularities that we can rely on. As it turns out, many things are unreliable in the case of these digitized manuscripts. The images have different sizes. There is no guarantee of a flap. The number of points needed to define the hull is different for each photo. The flap might be left or right. The codex might not be exactly straight in the photo. Sometimes the tip is cut out of the picture, and so on. Sometimes, you think you can exploit a regularity. But then as you cycle through a few manuscripts, you come to one with an exception and the computer does something unexpected or it outright crashes.

After several unsuccessful attempts, I landed on the strategy to find the tip of the flap by looking only at points of the hull that were within the middle 30% height-wise. Since we do a fair number of mathematical operations in this function, we need to rely on the *NumPy* package. Moreover, for that to work best we need to convert our list to NumPy arrays. We establish a bandwidth of Y values and make an array that has as many zeros as there are coordinates. This array, *decideMiddlePoints*, will be filled with the value 1 in every place where the Y value of a coordinate is within the bandwidth. This will be useful because we can exploit the fact that multiplying by one preserves the value and multiplying by zero is zero. Thus, we can multiply the array with all the X values with this *decideMiddlePoints*. We do the same for the array with the Y values and, then, we have arrays that only have values for points that fall in that bandwidth, while still maintaining the original length. The rest are all zeros since multiplying by zero is zero. If you are already starting to lose your grip on what we are attempting to do here, a good way to once again understand what is happening is to run the script on an example image and print out all kinds of variables at various points (and include an *exit*() command right after to stop the script) to see how the variables are filled with values. Even better is to tinker with the formulas and see what kind of effect that has on the values in the variables.

Using a *for*-loop, we cycle through all values of the array that contains the Y-values of the corners of the hull, and we check to see if that Y value is in the middle 30% bandwidth. If so, we place a 1 in that position in the *decideMiddlePoints*. Subsequently, we perform another check. If there is another point that is less than (or equal to) 5 pixels away X-wise, and the corresponding Y value for that X value is within 20% of the top of the image or of the bottom, we know that wherever this middle point is, it cannot be a part of the flap. It could only

```
    upperAngles.append( np.rad2deg(np.arccos((A**2 + C - B**2) /
(2 * A * np.sqrt(C)))) )

  lowerAngles = []

  for pointNumber in range(0, len(lowerEdgeX)):
    A = abs(tipLowerX - lowerEdgeX[pointNumber])
    #A is 0 which is causing the problem. Not allowed to
divide by 0.
    B = abs(tipLowerY - lowerEdgeY[pointNumber])
    C = A**2 + B**2
    lowerAngles.append(np.rad2deg(np.arccos((A ** 2 + C - B
** 2) / (2 * A * np.sqrt(C)))))

  # Entire angle, rounded to zero decimals (would give false
sense of accuracy otherwise) is:
  if not upperAngles or not lowerAngles:
    print("Error calculating angle!")
    return (None, None, None, None, None, None, None)
  else:
    finalAngle = int(np.rint( np.average(np.
asarray(lowerAngles)) + np.average(np.asarray(upperAngles)) ))
    return(finalAngle, tipX, tipY, lowerEdgeX[-1],
lowerEdgeY[-1], upperEdgeX[-1], upperEdgeY[-1])

# -------------------------------------------------------------

def Display_image(imageNumber, original, analyzed,
hullContourX, hullContourY):
# Convert the analyzed result from BW to Color so that we can
draw in color over it
  analyzedConverted = cv2.cvtColor(analyzed,cv2.
COLOR_GRAY2BGR)
  imageHeight, imageWidth, _ = analyzedConverted.shape

# Do you want to see the original or the processed image?
  global imageAppearance
  if imageAppearance == "BW":
    imageForDisplay = analyzedConverted
  elif imageAppearance == "Color":
```

be part of a non-flap side, or perhaps the computer was unable to distinguish the codex from the background. So, we not only rule out that middle point, we also rule out all points on that side of the image. With side I mean here the left from the middle or right from the middle. If we did indeed find such a point adjacent to a middle point, then we just make *decideMiddlePoints* zero for all points, middle points or not, on that side of the image, and we no longer need to look for adjacent points; hence, the *break* command, which prematurely ends the *for*-loop.

We should be left with an array *decideMiddlePoints* that has the value 1 in each place where the corner points of the hull have something to do with the tip of the flap. If *decideMiddlePoints* only has zeros, the computer could not detect a flap. If there are some ones, we now split up the corner points of the hull into those that have to do with the tip of the flap and those that do not. We can do so very easily in Python by using a condition within the index. Written out, line 112 says the following: get the Y values of all points that are close to the tip of the flap by taking only those Y values of the points that define the hull around the white blob (which is the codex) where the corresponding value of *decideMiddlePoints* is not zero.

We now have one or more candidates for the tip of the flap. Since the tip is the point that sticks out the most, to find the tip can now be done in two steps. First, we check on which side of the image the tip-points are. If they are on the left of the middle, the tip-point must be the point which has the smallest X-value, which can be found by *.min*(), otherwise we can use *.max*(). If the flap extends outside the image, the tip itself is not visible and two points would be found, for which both points either have X = 0 or X = width of image. The Y value can then be found as an average of the two points. This *tipY* is only used for visualization and not for angle calculation.

The next block of code, I am not entirely sure about. I encountered some cases in which the computer produced a strange result or no result at all but an error, and it seemed I could step over those cases by including a check if no other point of the hull is right next to tip, within 4 pixels distance. This check is only done on points of the hull that are out of the 30% bandwidth. So instead of trying to find the angle in these cases with more effort, I simply discard them as cases in which the computer could not find the angle. This strategy relies on enough cases to succeed that a slightly bigger loss is of no concern. If your use case cannot afford such a loss, you would have to invest more time to research what is going on in these unusual cases and what can be done to measure the angle nonetheless.

```
    imageForDisplay = original
  else:
    print("Variable imageAppearance must be set to BW or 
Color!")
    exit()

  #Drawing all points of the outer, reduced hull
  for correctYvalue in range(0,len(hullContourX)):
      xpos = hullContourX[correctYvalue]
      ypos = hullContourY[correctYvalue]
      cv2.circle(imageForDisplay, (xpos, ypos), 3, (255, 0, 0), 
-1)

  #Calling Angle finding function
  message, tipX, tipY, lowerEdgeX, lowerEdgeY, upperEdgeX, 
upperEdgeY = Find_angle(hullContourX,hullContourY,imageWidth, 
imageHeight)
  if message == None:
    message = "No Flap Detectable!"
  else:
    message = "Angle of flap is {0} degrees.".format(message)
    cv2.circle(imageForDisplay, (tipX, tipY), 9, (0, 50, 
250), -1)
    cv2.circle(imageForDisplay, (lowerEdgeX, lowerEdgeY), 6, 
(0, 250, 250), -1)
    cv2.circle(imageForDisplay, (upperEdgeX, upperEdgeY), 6, 
(0, 250, 250), -1)

# Putting text and showing image
  cv2.putText(imageForDisplay,str(message),(int(imageWidth/2 - 
139), int(imageHeight/2 + 1)),cv2.FONT_HERSHEY_DUPLEX, 1, (0, 0, 
0), 1)
  cv2.putText(imageForDisplay,str(message),(int(imageWidth/2 - 
140), int(imageHeight/2)),cv2.FONT_HERSHEY_DUPLEX, 1, (0, 0, 
0), 1)
  cv2.putText(imageForDisplay,str(message),(int(imageWidth/2 - 
141), int(imageHeight/2 -1)),cv2.FONT_HERSHEY_DUPLEX, 1, 
(120, 0, 220), 1)
  cv2.imshow("{0}".format(imageNumber), imageForDisplay)
```

6.5 *The Mathematics behind Finding the Angle*

The rest of the code relies on the idea that the angle of the flap can be found if you add up the angle of the line that goes horizontally straight through the tip and the line from the tip along the upper edge of the flap, and the angle with the line from the tip along the lower edge of the flap. Let us do some back-of-the-envelope mathematics to see how that works. From there, we can write in plain English a pseudo-code of what we need to do, and then we can convert that into Python and NumPy code.

Following the figure below, we can see how we can split the entire angle of the flap into two angles, one top, one bottom. Further, according to the theories of similar triangles, it does not matter where we find the points to construct the angle. If part of the flap is cut off from view, we simply calculate not from the tip but from the first visible point, and we use a horizontal line through that point. Similarly, again, according to what we know of similar triangles, we do not need to know the actual topmost and bottommost corner point of the flap, as any point along the edge of the flap will do. Last, if the photo of the manuscript is a bit skewed, this does not matter, since the loss of angle on one side is made up by what is gained on the other side.

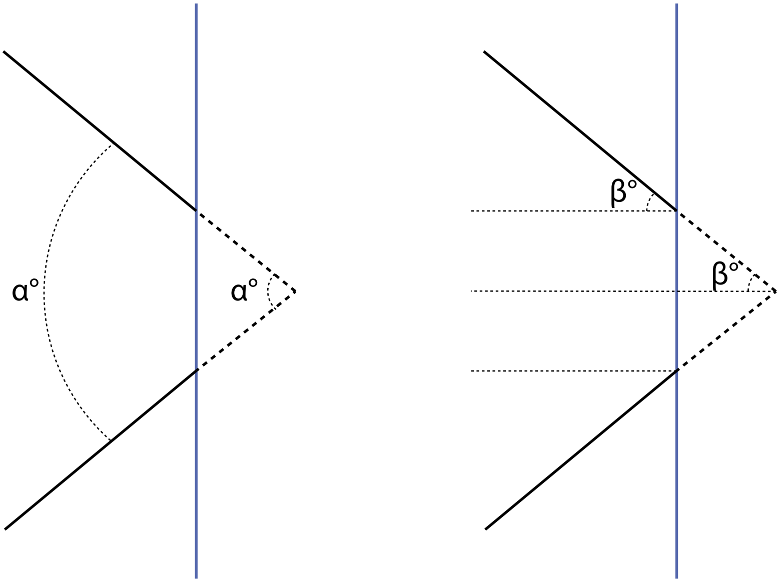


FIGURE 7.1 Proof that we can measure the top and bottom angle to obtain the entire angle of the flap

```
# Awaiting keyboard input.
# The key ‘q’ will show the manuscript with call number one
less than current.
# The key ‘w’ will show MS with call nr. one more.
# Any other key exits the program.
 key = cv2.waitKey(0) & 0xFF

 if key == ord(“q”):
  cv2.destroyAllWindows()
  Check_image_readable(imageNumber, False)
 elif key == ord(“w”):
  cv2.destroyAllWindows()
  Check_image_readable(imageNumber, True)
 elif key == ord(“e”):
  if imageAppearance == “Color”:
   imageAppearance = “BW”
   Display_image(imageNumber, original, analyzed,
hullContourX, hullContourY)
  else:
   imageAppearance = “Color”
   Display_image(imageNumber, original, analyzed,
hullContourX, hullContourY)
 else:
  cv2.destroyAllWindows()

# -------------------------------------------------------------

# Execute the code by starting the first function.
Check_image_readable(imageStartingNumber-1,
imageStartingDirection)
```

We have enough information if we can find the coordinates for two points along the (upper or bottom) edge of the flap. By taking the absolute value of subtracting their Y values, we get the 'height' of the triangle. By taking the absolute value of subtracting their X values, we get the 'width' of the triangle. Now, we have constructed a right triangle (see the figure) for which the theorem of Pythagoras counts, which allows us to calculate the length of the flap side of the triangle.

With the length of all three sides known, we can use the law of cosines which works on any triangle. This law says that the following formula is true, with *a*, *b*, and *c* the length of the edges and *y* the angle of a corner: $c^2 = a^2 + b^2 – 2ab \cos y$. This formula expresses the length of one side into the lengths of the other sides and the cosine of the angle of the opposite corner. We can flip the formula to express the cosine of the angle into the lengths of the triangle as such: $\cos y = (a^2 + b^2 – c^2) / 2ab$. To get to the angle *y*, we then need to take the inverse of the cosine, which is called arccos. We do not need to know much more about what *cos* and *arccos* really are since we will let NumPy do the calculations for us.

In other words, we need to find any two points along the upper edge of the flap to calculate the upper angle, do the same for the lower edge of the flap. Then, by putting them together, we get the angle of the flap. To be sure and accurate about this, we shall tell the computer to find a couple of points along each edge and take the average of the angles. We will likely get an answer with many decimals, but this is a false sense of accuracy. To do justice to our method, we shall round off the answer to whole degrees, as I think no greater degree of accuracy may be expected.

Getting back to our code, we first establish the tip points, which may be just one if the tip of the flap is visible on the image. But it may be two if the tip falls out of view. As we proved above, the two points can stand in for the tip while calculating the angle of the flap.

We need to distinguish two cases: if the flap is left or if the flap is right. If the flap is left, we distinguish two further cases: the points along the upper edge and the points along the lower edge. Looking at the upper edge case, we find the points in two steps: first, we find all points on the upper side of the image, then among those points we find all points that are within 20% of the left side of the image. By including that condition within the index, we can write our code in a very compact manner. Notice that the order is important here: we reduce our coordinates to the upper half by first doing so for X values, then Y; for the reduction to the left quadrant we first do so for Y values, then X. Otherwise, the index numbers would not be accurate. As a last act, we define two functions we will be using later on. By defining them here, we keep the number of lines of code to a minimum, since the same functions are defined slightly different in the case of the flap being on the right. If we do not define them as functions here, we would have to include complicated *if*-evaluations later, which would add more lines of code and would have made our code much less readable. The functions for the flap being on the left look into the collection of points in the upper (or lower) quadrant and give back the index of the point which is mostly to the left. We also perform some actions to get that index number in the shape we prefer.

With the knowledge of the coordinates of the point that stands in for the tip for the top and bottom edge of the flap, we now want to know the coordinates for some points along the edge. We only have a collection of points in the top (and bottom) quadrant, but do not know how many and in what order they are stored. We want three pieces of information and it will be better to separate them out. These are the number of the point, its X coordinate, and its Y coordinate. The number is something we assign ourselves. A maximum number is set as four at the beginning and can be adjusted by hand. The *for*-loop takes care of this. Even though we might want four points along the edge, there may only be three, two, or just one. That is what the *if*-statement takes care of. We then find the index for the point with the lowest X value (if the flap is on the left) and use that index value to add the appropriate X and Y value to their respective lists. Then, we delete that point so that the next time *Get_upper_point_number* is called, it will not return the same point.

Once we have the X and Y coordinates of some points along the upper and lower edges of the flap, we can calculate the lengths of the triangle they make with the tip of the flap (or the point that stands in for the tip). This is done in a *for*-loop as it needs to be done separately for each point. Since the angle calculation depends on dividing by *A* (the length X-wise), and dividing by zero is not allowed, we incorporate a safety check and use a *try*-statement for the formula of the cosine law. Additionally, NumPy gives back an answer in radians, not degrees, for its *arccos*-function. So, we add a *rad2deg*-function to ensure the answer is provided in degrees. We end up with two lists: a couple of angles for the upper part of the flap, and a couple of angles for the lower part of the flap. In the variable *finalAngle*, we add up the averages of each and round it off to whole numbers and ensure we are left with only an integer, not a list. We have just programmed the computer to give us the angle of the flap!

7 Step 3: Running the Script over Large Numbers

In the script of this step, I have combined the previous two scripts into one and made it functionally different in the sense that we no longer want a visual representation of the angle, but want to have the computer analyze hundreds, possibly thousands, of manuscripts and output all of that data together. Instead of not giving the angle back, we actually give some information back, so that we can quickly see later how many manuscripts stopped working at which point in the code.

Most of the code is the same as before, with minor adjustments. We shall only discuss the important differences. At the beginning, we see an object defined with curly brackets, Python's way of saying this is a dictionary. Dictionaries are a lot like lists in the sense that you can stuff an arbitrary number of values in it. However, with the difference that everything you put in it, you must place them in pairs: the first value of the pair is the identifier, the keyword, the second word is the actual value. Thus, it works much like a real dictionary: for every keyword, there is a definition. These keywords must be unique, because if you wish to add something to the dictionary but use a previously used keyword, you do not add it but simply overwrite the value of the previous definition. Dictionaries are designed as though the order in which the entries are stored does not matter, and that is how they should be used. We will use a dictionary to save all the results and then write the dictionary to disk.

The function *Initiate_angle_finding* handles the whole process of analyzing an image, calling all other functions, and handling their responses. At the bottom of our code, we loop that function to cover all the manuscripts we want. We can either do so by looping through the number in which the filename ends, or simply by commanding the computer to inspect a certain folder and analyze all PDF files. This will generate a random order in which these files are analyzed and so it is useful to implement some *print* command to know the script is still running. This part of the script should be adapted if you have JPG files by uncommenting lines 71 and 72. If you work with a more complicated folder structure, you will have to find some regularity that you can exploit through a *for*-loop.

At the very end, a CSV file is created and saved on the hard disk. A CSV file is easy to import in, for instance, Excel or import back into Python. There is a *csv* package for Python which has specific functions to write all entries of a dictionary as a pair on a row.

```
# L.W. Cornelis van Lit, O.P. (c) 2017-2018
# Finding the angle of the flap of an Islamic manuscript

# -----------------------------------------------------------

# Dependencies
import cv2
import numpy as np
import io
from PyPDF2 import PdfFileMerger
import csv
import os
import time

#For analytical purpose only, a timer is set
start_time = time.time()

# A dictionary is created to catch the results
results = {}

# A string defining a JPG always starts and ends with these
values
startmark = b”\xff\xd8”
endmark = b”\xff\xd9”

#User variables
maximumEdgePoints = 4
kernelbig = np.ones((10, 10), np.uint8)

# -----------------------------------------------------------

def Initiate_angle_finding(pathOfFiles, nameOfPDF, pageOfPDF):

# Check if file exists and is sound. Change “.pdf” to .jpg”
to use jpgs instead of pdfs and uncomment lines 67 and 68.
  if not Accessing_file(pathOfFiles + nameOfPDF + “.pdf”, ‘rb’):
    results[nameOfPDF]=“File inaccessible.”

# Check if image on specific page can be extracted.
  else:
    extractedImage = Extract_image_from_PDF(pathOfFiles,
nameOfPDF, pageOfPDF)
    if not type(extractedImage) == np.ndarray:
      results[nameOfPDF] = “Image inaccessible.”

# Perform analysis on image.
    else:
      hullContourX, hullContourY, imageWidth, imageHeight =
Analyze_image(extractedImage)
      if not hullContourX:
        results[nameOfPDF] = “Cannot analyze image.”
      else:
        angleInfo = Find_angle(hullContourX, hullContourY,
imageWidth, imageHeight)
        results[nameOfPDF] = angleInfo

# ------------------------------------------------------------

# Check if a file exists and is accessible.
def Accessing_file(filepath, mode):
  try:
    f = open(filepath, mode)
    f.close()
  except:
    #File unreadable.
    return False
  #File readable.
  return True

# ------------------------------------------------------------
# This function takes one PDF file and extracts one page from
it, to save as JPG.
# Only to be used on PDFs in which each page is only an image.

def Extract_image_from_PDF(pathOfFiles, nameOfPDF, pageOfPDF):

#Uncomment these two lines of code if files are already images:
# img = cv2.imread(pathOfFiles + nameOfPDF + “.jpg”)
# return img
# File exists, so prepare the virtual containers.
  merger = PdfFileMerger()
  virtualpdf = io.BytesIO()

# Opens PDF and takes out specific page
  with open( pathOfFiles + nameOfPDF + “.pdf”, “rb”) as
sourcePDF:
    try:
      merger.append(fileobj=sourcePDF, pages=(pageOfPDF - 1,
pageOfPDF))
      merger.write(virtualpdf)
    except:
      return
  merger.close()

# Read the desired page simply in its string of values
  pdf = virtualpdf.getvalue()
  virtualpdf.close()

# Find the start and end of the string defining the JPG
  jpgstart = pdf.find(startmark, 0)
  jpgend = pdf.find(endmark, jpgstart)

# Reading out the entire string defining the JPG
  jpgstring = pdf[jpgstart:jpgend]

# Preparing the string to be read by OpenCV
  jpgstring2 = np.fromstring(jpgstring, np.uint8)

# Turn into OpenCV-readable image
  image = cv2.imdecode(jpgstring2, 1)

# Give back OpenCV-readable image
  return image

# -------------------------------------------------------------

#This function analyses the image.
def Analyze_image(image):
  gray = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
  ret, thresh1 = cv2.threshold(gray, 180, 255, cv2.
THRESH_BINARY_INV)
  # Twice scaling is part of accessing the angle
  scaled = cv2.resize(thresh1, (0, 0), fx=0.8, fy=0.8)
  openimg = cv2.morphologyEx(scaled, cv2.MORPH_OPEN, kernelbig)
  ret, thresh2 = cv2.threshold(openimg, 1, 255, cv2.
THRESH_BINARY)
  scaledagain = cv2.resize(thresh2, (0, 0), fx=0.5, fy=0.5)

  imageHeight, imageWidth = scaledagain.shape

  _, contours, _ = cv2.findContours(scaledagain, cv2.RETR_
EXTERNAL, cv2.CHAIN_APPROX_SIMPLE)
  try:
    mainContour = max(contours, key=len)
  except:
    return (None,None,None,None)

  hull = cv2.convexHull(mainContour, returnPoints=False)

  hullContourX = []
  hullContourY = []
  for correctYvalue in hull:
    hullContourX.append(mainContour[correctYvalue[0]][0][0])
    hullContourY.append(mainContour[correctYvalue[0]][0][1])

  mainContourX = []
  mainContourY = []

  mainContourTotalX = 0
  mainContourTotalY = 0
  for correctYvalue in range(0,len(mainContour)):
    mainContourX.append(mainContour[correctYvalue][0][0])
    mainContourY.append(mainContour[correctYvalue][0][1])
    mainContourTotalX = mainContourTotalX +
mainContour[correctYvalue][0][0]
    mainContourTotalY = mainContourTotalY +
mainContour[correctYvalue][0][1]

  return hullContourX, hullContourY, imageWidth, imageHeight
```

```
# -----------------------------------------------------------

def Find_angle(hullContourX, hullContourY, imageWidth,
imageHeight):
  # Using NumPy requires arrays
  reducedXArray = np.array(hullContourX)
  reducedYArray = np.array(hullContourY)

  # Using NumPy to get index numbers of all Y values within
15% of middle
  reducedYLength = len(hullContourY)
  varianceY = imageHeight * 0.15
  minimumY = imageHeight / 2 - varianceY
  maximumY = imageHeight / 2 + varianceY
  decideMiddlePoints = np.zeros(reducedYLength, dtype=np.int)

  for correctYvalue in range(0, reducedYLength):
    if minimumY <= hullContourY[correctYvalue] <= maximumY:
      np.put(decideMiddlePoints, correctYvalue, 1)
      xVal = hullContourX[correctYvalue]

      # Ruling out spurious points on non-flap side by
checking if nearby points in terms of x-value are near edge
in terms of y-value (i.e. in a corner)
      # Also ruling out flaps with distortions, i.e. deleting
all middle points on the left/right side of the image where
there is a point which meets above requirement.
      for otherXvalues in range(0, reducedYLength):
        if (xVal - 5 <= hullContourX[otherXvalues] <= xVal
+ 5) and (hullContourY[otherXvalues] < imageHeight * 0.2 or
hullContourY[otherXvalues] > imageHeight * 0.8):
          if hullContourX[otherXvalues] <= imageWidth/2:
            np.put(decideMiddlePoints,np.where( reducedXArray
< imageWidth / 2), 0)
          else:
            np.put(decideMiddlePoints, np.where(reducedXArray
> imageWidth / 2), 0)
          break
  # Only proceed if a flap can be found.
  # Get X,Y value of tip-point, making sure there are some
points that could be the tip

  if not np.count_nonzero(decideMiddlePoints) == 0:
    # Get X,Y value of only tip-points
    tipOfFlapYArray = reducedYArray[decideMiddlePoints != 0]
    tipOfFlapXArray = reducedXArray[decideMiddlePoints != 0]
    restOfHullXArray = reducedXArray[decideMiddlePoints == 0]
    restOfHullYArray = reducedYArray[decideMiddlePoints == 0]

    # tipX needs to be decided, left or right
    # Is the tip on the left?
    if np.average(tipOfFlapXArray) < imageWidth / 2:
      # Get X,Y value of furthest tip-point
      reducedXMinimumArray = np.where(tipOfFlapXArray ==
tipOfFlapXArray.min())
      reducedXMinimumList = reducedXMinimumArray[0].tolist()
      tipX = int(tipOfFlapXArray.min())

    # Is the tip on the right?
    elif np.average(tipOfFlapXArray) > imageWidth / 2:
      reducedXMinimumArray = np.where(tipOfFlapXArray ==
tipOfFlapXArray.max())
      reducedXMinimumList = reducedXMinimumArray[0].tolist()
      tipX = int(tipOfFlapXArray.max())

    else:
      return (“Middle points undetermined.”)
  else:
    if len(hullContourX) > 4:
      return (“Cannot find flap.”)
    else:
      return (“No flap.”)

  # Check if hull is formed correctly
  if (abs(restOfHullXArray - tipX) < 4).any():
    return (“Some point of hull too close to X value of tip.
Hull incorrectly identified.”)
```

```
 # Now find points along the edge and calculate angle

 tipUpperX = tipX
 tipUpperY = np.min(tipOfFlapYArray[reducedXMinimumList])
 tipLowerX = tipX
 tipLowerY = np.max(tipOfFlapYArray[reducedXMinimumList])

 #For left flap we need to look at minimal X values, for
right flap maximal X values.
 if tipX < imageWidth / 2:

  # Get all XY that are in the upper quadrant
  upperRestOfHullXArray = restOfHullXArray[restOfHullYArray
< imageHeight / 2]
  upperRestOfHullYArray = restOfHullYArray[restOfHullYArray
< imageHeight / 2]

  upperRestOfHullYArray = upperRestOfHullYArray[upperRestOf
HullXArray < imageWidth * 0.2]
  upperRestOfHullXArray = upperRestOfHullXArray[upperRestOf
HullXArray < imageWidth * 0.2]

  # Get all XY that are in the lower quadrant
  lowerRestOfHullXArray = restOfHullXArray[restOfHullYArray
> imageHeight / 2]
  lowerRestOfHullYArray = restOfHullYArray[restOfHullYArray
> imageHeight / 2]

  lowerRestOfHullYArray = lowerRestOfHullYArray[lowerRestOf
HullXArray < imageWidth * 0.2]
  lowerRestOfHullXArray = lowerRestOfHullXArray[lowerRestOf
HullXArray < imageWidth * 0.2]

  def GetUpperPointNumber():
   return((np.where(upperRestOfHullXArray ==
upperRestOfHullXArray.min())[0]).tolist())

  def GetLowerPointNumber():
      return((np.where(lowerRestOfHullXArray ==
lowerRestOfHullXArray.min())[0]).tolist())
  else:
    # Get all XY that are in the upper quadrant
    upperRestOfHullXArray = restOfHullXArray[restOfHullYArray <
imageHeight / 2]
    upperRestOfHullYArray = restOfHullYArray[restOfHullYArray
< imageHeight / 2]

    upperRestOfHullYArray = upperRestOfHullYArray[upperRestOf
HullXArray > imageWidth * 0.8]
    upperRestOfHullXArray = upperRestOfHullXArray[upperRestOf
HullXArray > imageWidth * 0.8]

    # Get all XY that are in the lower quadrant
    lowerRestOfHullXArray = restOfHullXArray[restOfHullYArray >
imageHeight / 2]
    lowerRestOfHullYArray = restOfHullYArray[restOfHullYArray >
imageHeight / 2]

    lowerRestOfHullYArray = lowerRestOfHullYArray[lowerRestOf
HullXArray > imageWidth * 0.8]
    lowerRestOfHullXArray = lowerRestOfHullXArray[lowerRestOf
HullXArray > imageWidth * 0.8]

    def GetUpperPointNumber():
      return((np.where(upperRestOfHullXArray ==
upperRestOfHullXArray.max())[0]).tolist())
    def GetLowerPointNumber():
      return((np.where(lowerRestOfHullXArray ==
lowerRestOfHullXArray.max())[0]).tolist())

  # Establishing upper edge points
  upperEdgePoint = []
  upperEdgeX = []
  upperEdgeY = []
  for pointNumber in range(0, maximumEdgePoints):
    if pointNumber < len(upperRestOfHullXArray):
      upperEdgePoint.append( GetUpperPointNumber() )
      upperEdgeX.append( upperRestOfHullXArray[upperEdgePoint
[pointNumber][0]] )
      upperEdgeY.append(upperRestOfHullYArray[upperEdgePoint[
pointNumber][0]])
      upperRestOfHullXArray = np.delete(upperRestOfHullXArray,
upperEdgePoint[pointNumber])
      upperRestOfHullYArray = np.delete(upperRestOfHullYArray,
upperEdgePoint[pointNumber])

  # Establishing lower edge points
  lowerEdgePoint = []
  lowerEdgeX = []
  lowerEdgeY = []
  for pointNumber in range(0, maximumEdgePoints):
    if pointNumber < len(lowerRestOfHullXArray):
      lowerEdgePoint.append( GetLowerPointNumber() )
      lowerEdgeX.append(lowerRestOfHullXArray[lowerEdgePoint[
pointNumber][0]])
      lowerEdgeY.append(lowerRestOfHullYArray[lowerEdgePoint[
pointNumber][0]])
      lowerRestOfHullXArray = np.delete(lowerRestOfHullXArr
ay, lowerEdgePoint[pointNumber])
      lowerRestOfHullYArray = np.delete(lowerRestOfHullYArr
ay, lowerEdgePoint[pointNumber])

  # !!! Note that upper/lowerRestOfHull are now empty !!!

  # Calculate degree of angle based on lower edge points
  # First apply Theorem of Pythogaras to find all lengths,
with A = horizontal, B = vertical, C = diagonal

  upperAngles = []
  for pointNumber in range(0,len(upperEdgeX)):
    # First apply Theorem of Pythogaras to find all lengths,
with A = horizontal, B = vertical, C = diagonal
    A = abs(tipUpperX - upperEdgeX[pointNumber])
    B = abs(tipUpperY - upperEdgeY[pointNumber])
    C = A**2 + B**2
    # Then use Law of Cosine to find angle. Answer is returned
in Radians so it needs to be transposed to Degrees for human
readability.
    try:
      upperAngles.append( np.rad2deg(np.arccos((A**2 + C -
B**2) / (2 * A * np.sqrt(C)))) )
    except:
      return (“Upper points incorrectly identified.”)

  lowerAngles = []
  for pointNumber in range(0, len(lowerEdgeX)):
    A = abs(tipLowerX - lowerEdgeX[pointNumber])
    #A is 0 which is causing the problem. Not allowed to divide
by 0.
    B = abs(tipLowerY - lowerEdgeY[pointNumber])
    C = A**2 + B**2
    try:
      lowerAngles.append(np.rad2deg(np.arccos((A ** 2 + C - B
** 2) / (2 * A * np.sqrt(C)))))
    except:
      return (“Upper points incorrectly identified.”)

  # Entire angle, rounded to zero decimals (would give false
sense of accuracy otherwise) is:
  if not upperAngles or not lowerAngles:
    return (“Cannot make out angle.”)
  else:
    finalAngle = int(np.rint( np.average(np.
asarray(lowerAngles)) + np.average(np.asarray(upperAngles)) ))
    return(finalAngle)

# ------------------------------------------------------------

collectionName = “Nuruosmaniye-1-500”
directoryPath = “/Volumes/ManuscriptsHD/” + collectionName +“/”
directory = os.fsencode(directoryPath)

for file in os.listdir(directory):
  filename = os.fsdecode(file)
  if filename.endswith(“.pdf”):
    if not filename.endswith(“_text.pdf”):
      filename = filename[:-4]
      print(“Working on manuscript “ + filename)
      Initiate_angle_finding(directoryPath,filename,1)

#We can also loop the function over the number in which the
filename ends by using these lines of code:
#collectionName = “NURUOSMANIYE”
#startNumber = 1
#finishNumber = 362
#for i in range(startNumber,finishNumber+1):
# print(“Working on manuscript ” + collectionName + “ ” +
str(i))
# Initiate_angle_finding(“/Volumes/ManuscriptsHD/Nuruosmaniye/”
,collectionName+str(i), 3)

# Save results as CSV
with open(collectionName+“.csv”, “w”) as output:
  writeToCSV = csv.writer(output)
  writeToCSV.writerows(results.items())

print(“Finished in %s seconds.” % (time.time() - start_time))
```

8 Results

Arguably, what we created so far is not a final research result in its own right but only a stepping stone. More scripts could be developed to analyze the data we generated; for example, we could clean our data by doing our best to get rid of false positives. This could be achieved with a script to display the lowest and highest results to visually inspect whether they have correctly been calculated or whether the computer could not deduce the angle correctly. This could simply include a user interface where, with the keys Y and N, we could accept or decline the results; and if we decline, the result would be deleted from the CSV file. There are other issues that might muddy our data. For example, maybe we want to delete those codices which are not in their original binding, and so we would need to find further regularities to exploit. One possibility is to analyze the entire shape of a codex and compare it to the shapes of all the other codices, and if there is a (near)perfect match, we could flag it as a possible rebind.

Perhaps our input comes from digitized microfilms, and we might need to alter the algorithms to analyze those particular kinds of images.

After data cleaning, we could develop scripts to dynamically load all results and visualize in graphs the statistics that they produce, for instance, by using the *MatPlotLib* package. Even better, we could try to obtain other metadata on the manuscripts, for instance their date and origin, to try to relate the angle to specific eras and places. On the other hand, we could see the angle detection as only a first step to detect other features of the manuscript that we might be able to pinpoint if we can verify that certain points belong to the tip and the edges of the flap. Here is not the place to go into those things. Let us instead go over the final results.

TABLE 7.1 Number of manuscripts per outcome

Hull incorrectly identified	39
No flap	214
Image inaccessible	38
File inaccessible	83
Cannot make out angle	10
Cannot find flap	532
Angles	1,084
Total	**2,000**

In total, I processed approximately two-thousand manuscripts, which took me about half a day. The number of angles calculated seems disappointingly low given the total. However, this was expected as many of the files that I had access to did not have a photo of the cover of the codex; and for some that did, they did not have that photo on the first page. A logical step would be to look into the 'cannot find flap' category and see if it consists mostly of false negatives. If so, perhaps the code could be adjusted to collect more angles. I did not attempt this at this point. Checking parts of the processed images by hand yielded a success rate of 98%. The two percent false positives had to do with extra objects remaining in the background which OpenCV analyzed as part of the codex.

Of the manuscripts for which the angle could be determined, it became clear that there is a pretty significant concentration around certain angles. The number of codices that were identified as having a flap with an angle below 150° was 8, accounting for 0.7% of the total. Codices with a flap above 170° were 7 in total, or 0.6%. That means that 98.7% of the codices has a flap with an angle between 150° and 170°. Further, degrees 156 through 160 are the only ones with a count above a hundred, and 158° sticks out high above the rest, with 165 manuscripts. The diagram at the end of this chapter shows the number of manuscripts with a flap that has an angle for a certain degree. The angle can obviously only be between 0° and 180°. Among the ones that are counted below 150° and above 170°, we should assume there are some false positives. Nevertheless, I do not think we need to clean our data as the vast number of manuscripts convincingly indicate the matter here, namely that the flap (*lisān*) of an Islamic manuscript will normally make an angle of about 156° to 160° degrees, as this narrow bandwidth already accounts for 60% of all manuscripts. That being said, clearly the flap of codices was not produced with a standardized tool, as the other 40% falls outside of this bandwidth and we do see considerable variety.

With this chapter, then, I have shown that codicology too can benefit from the mass digitization of manuscripts. Python and OpenCV are powerful and stable pieces of technology. Even if they will be superseded, the skills gained from using them will be easily transferrable to other programming languages. What is most important is not to know certain commands and functions by heart, but to have a general idea what a programming language or a package is capable of, to write out pseudo-code along those lines, and then to find out how to exactly implement this from a technical point of view. The technology might still throw you some curve balls. But workarounds often present themselves in online discussions among people who have previously dealt with a similar issue. As long as your code achieves what you want, is readable, and reasonably flexible, you do not need to worry about the rest. As a result, I maintain that virtually anybody can instruct a computer to perform fairly advanced analyses on images of manuscripts, automated over large amounts of data.

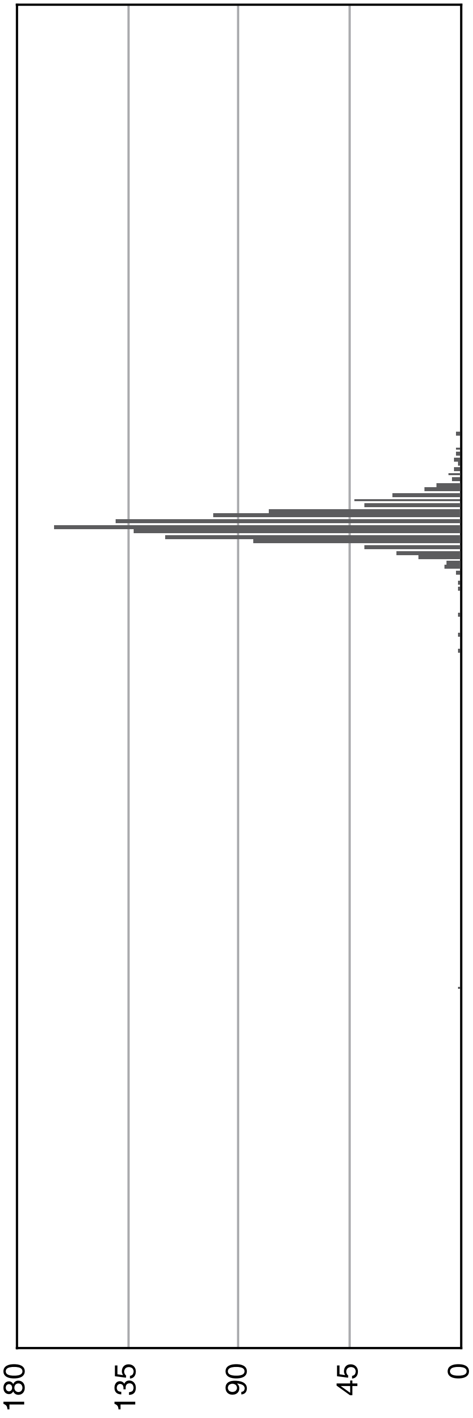


FIGURE 7.2 Bar chart of measured angles for a thousand manuscripts

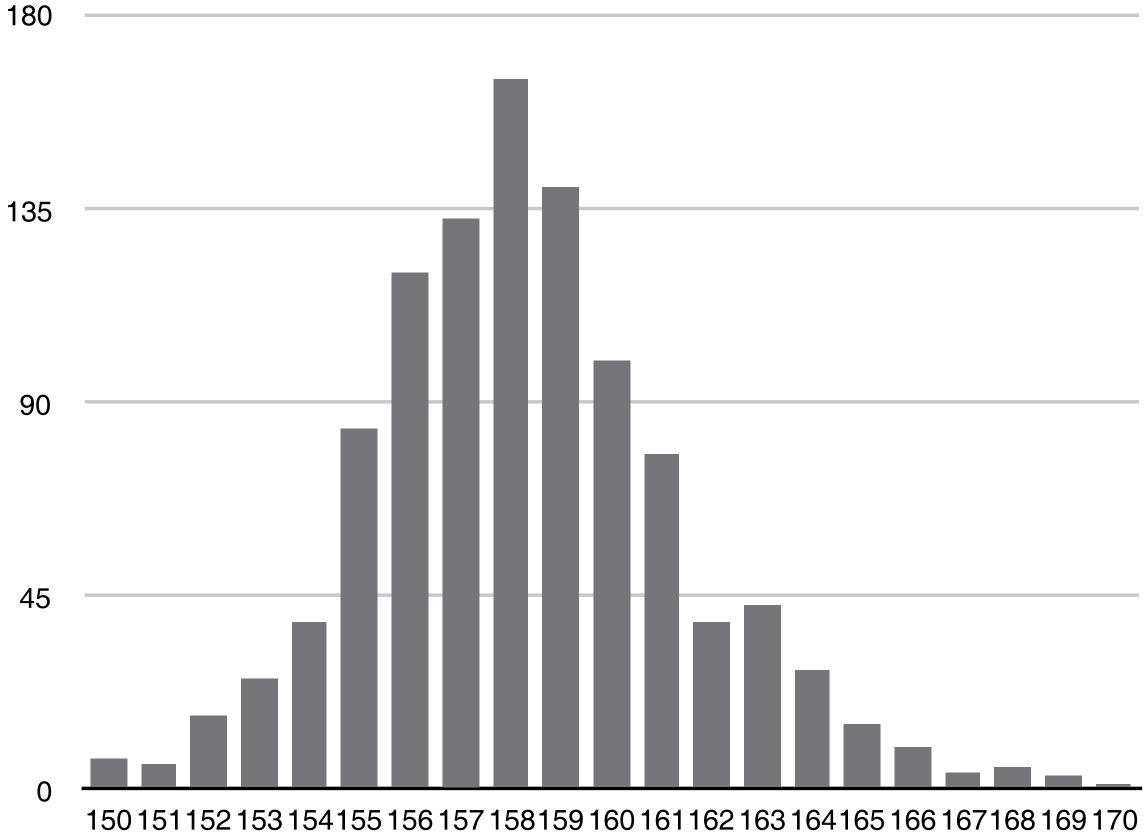


FIGURE 7.3 Close up showing most manuscripts have a flap with an angle of around 158°

CHAPTER 8

A Digital Orientalist

How can we hold the digital world in one hand and the manuscript world in the other? From 2013 onwards, all my activities involved in putting computers to the service of humanities studies were grouped under the project name 'The Digital Orientalist.' 'Digital' referred to anything researchers can achieve by means of computers, especially the seemingly futuristic and limitless potential of obtaining ever larger data sets and manipulating them in whatever way they think is meaningful. The term 'Orientalist' is, of course, not aiming at the Orientalism that Edward Said so profoundly dismantled[1] but to an archaic grouping of fields of studies, among them my own Islamic studies, which have the usage of ancient manuscripts in common, as their most important source for the study of civilizations to the East of Europe.[2] At first glance, the combination of these two terms seems to be an oxymoron. However, by now it has become more and more evident that the supposed contradiction between what is 'Digital' and what is 'Orientalist' is, in fact, a fusion that leads to a great increase in our research capabilities.

This state of affairs makes for exciting times for classical fields of the humanities. Our fields often involve the painstaking collection of large amounts of snippets of evidence from disparate and scattered sources as well as close reading of texts to tease out its many layers of meanings, allusions, cross-references, and intertextuality. Much of these activities can be done cheaper, faster, and better by using digitized materials in a digital workflow. We can give a computer the boring stuff that test our patience and abilities, such as remembering where we kept every note we made or noticing a similarity between a new note and one we made a long while ago. As for us, we only need to focus on the interesting stuff, such as deciding whether such a similarity is significant or not. In addition, computers add entirely new tools to our toolbox, such as image manipulation and distant reading. 'Classical' humanities will thrive because of 'digital' humanities.

[1] Said, E. *Orientalism*. New York: Pantheon Books, 1978.

[2] As still witnessed by learned societies such as the American Oriental Society or the Deutsche Morgenländische Gesellschaft, and also still present in the name of renowned institutions such as Chicago's Oriental Institute, London's School of African and Oriental Studies, and Paris's Institut national des langues et civilisations orientales.

© CORNELIS VAN LIT, 2020 | DOI:10.1163/9789004400351_010

This new wave comes with an important challenge, namely, that we will have to invest ourselves in making these digital tools our own. This is only logical since our fieldwork, or rather the soil of our fields, has changed dramatically. To get to our manuscripts, we no longer go to dusty libraries but turn to buzzing computers, and if we do go to a library, we often come back with digital photos. Not so long ago, it was not expected of classically trained students and scholars to have computer skills.[3] As it becomes more and more obvious that a 'Digital Orientalist' is not an oxymoron, it will become increasingly necessary and expected that all researchers possess some understanding of and skills in computer technology.

What has reading this book been good for? We have gone through extensive conceptual and theoretical discussions and worked our way through a diverse palette of technical and practical skills. The theoretical discussions attempted to foster an awareness of the changing nature of our fieldwork through which we can avoid mistakes that would arise from approaching our digital workflow with a manuscript or print mindset. Two big mistakes highlighted in this book are (1) to disregard the digital materiality of digitized manuscripts, and (2) to limit oneself to print-like publications.

Issues surrounding the first mistake are addressed in Chapters Two and Three. From Chapter Three, we learned that not all digitized manuscripts are created equally. In other chapters, we prepared ourselves for this by learning skills to change digitized manuscripts to a more usable shape. In Chapter Five, we found out that this itself can introduce differences. For instance, we noticed that using different methods of converting a file from one format to another will yield different outcomes. In Chapter Two, we found the right vocabulary to describe this change and difference and introduced ten notions that can help us describe the digitized manuscript as a file and its *Sitz im Leben*. From now on, when we use a digital surrogate in our research, we are able to notify the reader of this in our publications by including a sentence that says something along the lines of 'The digital images used in this work are $X_{\text{px}} \times Y_{\text{px}}$ showing N pages, at about Z kb. The cut is tight and color balance is fine.' The X and Y values are the pixel **dimensions** of the actual images, N is usually one or two depending if the image is of one **page or a page-spread**, Z is the **file size**, the **cut** is a description of how much is visible around the page, and the **color balance** is a judgment on whether the colors in the image are true to life.

The implications arising from the second mistake are not simply a plea for digital editions. As addressed at the end of Chapter Five, born-digital

[3] Babeu, A. *"Rome Wasn't Digitized in a Day": Building a Cyberinfrastructure for Digital Classics*. Washington: Council on Library and Information Resources, 2011, p. 139.

publications could be the right outcome of your project but are not necessarily the most useful ones. More importantly, the print-publication mindset can limit us even far earlier in the creative process and this would be a serious mistake. In Chapters Four and Seven, we saw that digital surrogates can be manipulated in ways virtually unimaginable in the print world. While the first of the two above-mentioned issues chartered the limits of using digital surrogates, this second one highlights the incredible potential. Concurrently, it should be noted that this potential does not impact our tried and tested methodologies and project management: older approaches are still valid, as we are merely superimposing specific computer-supported methods. This is important to note since there is a mindset that has portrayed digital humanities as possibly antagonistic towards the fundamental methodologies of classical humanities. This mindset encouraged the shift towards big-grant team projects. In this book, I have forcefully argued against this shift. Such projects want to split technical and scholarly tasks and assign them to different people. In my opinion, we ought to combine in our own selves the technical know-how and the field-specific expertise. This can take many shapes; instead of thinking of 'digital humanities' as in binary opposition to 'classical humanities,' I advocate for seeing digital humanities as a spectrum. Only if we manage to fill that entire spectrum can we make the most of computer technology, and then each of our humanities disciplines can gradually progress in building an ecosystem of standards, tools, and technologies, and possibly repositories, that we actually use. Of course, this does not exclude the possibility to work within or to start a big-grant project. In fact, having verifiable experience in both technical and field-specific aspects will make you a better candidate for that.

Looking over this book, you will have found very few technologies that were specifically developed for the humanities (TEI and IIIF are the exceptions: see Chapter Five.) There are many technologies from other industries that we can profitably adapt and use. In Chapter Four, we cherry-picked tools from the graphic design industry; in Chapter Six, we reaped the low-hanging fruits of the web development industry; in Chapter Seven, we made thankful use of the tools developed by the industry that needs automated object detection on video footage. The tech industry today has an extraordinary 'sharing is caring' mentality, which means that many powerful tools are free to use and free to learn. In particular, the world of web development is easy to penetrate with its very large pool of learning materials. The goal, then, of this book is not so much to outfit you with the tools you need but, rather, after you have acquired the foundations, to motivate you to keep learning. In each of the above-mentioned technologies, changes are bound to happen. But with proper knowledge of the foundations, adjusting and adapting new software or new programming libraries will only be natural, as you continue to incorporate into your personal toolbox the latest, most interesting technology that is relevant to you.

The specific examples used in the chapters can be merged into powerful combinations. For example, you can draw vector-based SVG-images using what you learned in Chapter Four, then mass-manipulate them with Python using Chapter Seven, then present them through an interface using Chapter Six, keeping in mind the standards discussed in Chapter Five. Even if you intend to remain a passive consumer of technology, there is great merit in looking over these practical chapters. This will yield a sense of mastery, of being in control and, by extension, it will instill a creative liberty providing you with the confidence to handle digitized manuscripts and do with them whatever you want—nothing is impossible. Moreover, you will know which technical limits have been reached and can, consequently, temper the exaggerated expectations that some may have.

What can we expect in the future? Here is a sweeping statement by Jerome McGann from 2005: "In the coming decades—the process has already begun—the entirety of our cultural inheritance will be transformed and re-edited in digital forms."[4] Our own fields and their unstoppable wave of digitizing manuscripts are proof that this is true, so much so that we are already leaving behind the era in which it is a novel request to ask for digital images of a manuscript. Instead, we can begin to transform our workflow into digital forms. An excellent sign of this is will be the inclusion of digital methods in our teaching. As this progresses, the work now known as 'digital humanities' will become increasingly integrated into every field of the humanities in their own particular ways. Its application will become normal, even expected, and so the epithet 'digital' will lose significance.[5] Again, we should look out for the moment when specific training in 'digital methods' quiets down and inclusion in regular courses becomes expected.[6] Digital humanities as a field surely still has a future of its own, perhaps similar to Statistics which is used throughout the sciences but knows its own specialists with their own discourse and manifests itself as a department in its own right here and there. Nonetheless, it is the institutional embedding of this field that may need the greatest change. Already now, DH

[4] McGann, J. "Information Technology and the Troubled Humanities." pp. 105–21 in *TEXT Technology* 14, no. 2 (2005), p. 109.

[5] Cf. Zorich, D.M. "Digital Humanities Centers: Loci for Digital Scholarship." pp. 70–78 in *Working Together or Apart: Promoting the Next Generation of Digital Scholarship*. Washington, D.C.: Council on Library and Information Resources, 2009, p. 77.

[6] Note that this book's chapters refer to 'Codicology', 'Paleography,' and 'Philology' and not 'Digital Codicology' and so forth.

centers are criticized for being "silos of activity and redundant resources,"[7] and the people therein are accused of being "inward focused."[8] These characterizations are especially painful since the most recurrent values mentioned in digital humanities literature talk about desilofication, reuse, and expansion of our horizons. Part of the problem, as I see it, is that DH centers are almost always financed by generating income through grants. Such grants can only be obtained by proposing big projects. Then, digital humanities is locked into doing big projects that invariably foster an in group–out group culture and produce technology that is created and stored at a center, to the detriment of the periphery. To counter this situation, decentralization will hopefully become a key theme; decentralization of things such as repositories, tools, and standards but also of labor such as programming and encoding. In short, decentralization of responsibility. This would foster a positive stance towards open source and open access and encourage everybody to contribute with small additions. We know that in a manuscript culture, readers cast votes by choosing texts they like to write out and make new copies of. This way, texts either become popular or die out. Similarly, faced with the new challenges presented by the digital world, we do not have to create the perfect end solution in a single effort, nor expect our solution to be universally accepted. We can, instead, use a collection of small tools and standards, swap out one thing for another if something else works better, and if we add something of our own, then share it back with the rest of the world. Gradually, but perceptibly, our fields will turn digital. It has never been a better time to be studying ancient manuscripts.

[7] Zorich, p. 71.

[8] Prescott, A. "Consumers, Creators or Commentators? Problems of Audience and Mission in the Digital Humanities." pp. 61–75 in *Arts & Humanities in Higher Education* 11, no. 1–2 (2011), p. 67.

Postscript. Among Digitized Manuscripts

Working in manuscript studies is an extraordinary grace. In my own field, Ignati Kratchkovsky (Игнатий Крачковский) wrote beautifully about this, in his *Among Arabic Manuscripts*; a memoir of his experience with manuscripts throughout his life (1883–1951). He writes about what tactile interaction can do, saying for example that "Many are the hands through which it passed in Africa, Asia and Europe before it came to rest on the shelves of the Manuscript Department."[1] He speaks of the highs "when some discovery will gleam like a tiny spark,"[2] and of the lows "bringing me often to the verge of despair and making me doubt my ability."[3] Kratchkovsky loves the manuscripts and the libraries they are in, but already in his time photographic surrogates were used. He laments its use in a passage that is worth quoting in full:

> This feeling of attraction aroused by a copy of the original is familiar to all who work on manuscripts, for in our generation one is often obliged to work on photostatic reproductions, something which was unknown to our predecessors, who always worked either on the originals or on copies made by hand. However skillfully made, these latter could not reproduce many of the details and from them one could learn only the contents of a work without actually feeling a "live" manuscript with all its unrepeatable individual traits.[4]

In his book, he works out what that feeling of attraction is by giving little anecdotes, or snapshots, of his interaction with manuscripts, libraries, and the people in and around them. Each snapshot sets a certain tone and gives one aspect of the multi-faceted experience of working with manuscripts. Kratchkovsky does this masterfully, especially by connecting stories in unexpected ways. What seemed an unimportant detail in one snapshot, takes a central role in another one.

I think it is important to say something about what that experience is, when working with digitized manuscripts. To do that, I wish to leave you with some of my own stories, inspired in style and content by Kratchkovsky.[5]

[1] Kratchkovsky, I.Y., *Among Arabic Manuscripts*, transl. T. Minorsky, Leiden: Brill (1953), p. 31.

[2] Kratchkovsky, p. 91.

[3] Kratchkovsky, p. 76.

[4] Kratchkovsky, p. 163.

[5] I am also indebted to another citizen of Saint Petersburg of around the same time, Nicholas Roerich (1874–1947).

Death by Digitized Manuscript

One library relevant to my research has reached near-legendary status in making a fuzz about providing access to its manuscripts. One library kept the sole known manuscript copy of a text I was desperately after for my doctoral research. I mean one and the same: Topkapi Palace Museum Library. The name explains itself thus: Topkapi is the name of a former Ottoman palace that was turned into a museum, of which the manuscripts and books division was called a library. Scholars two generations above me reminisced of the time they hand-copied manuscripts in the reading room of Topkapi, with pencil and paper, as any kind of photographic reproduction was not tolerated. Scholars one generation above me were dismayed that now even readers were no longer tolerated in the reading room. For my own generation, Topkapi was merely a concept, without reality. To get to the manuscript I tried my usual tricks, asking colleagues if they got digital photos from there and if so how, especially by engaging my network in Turkey. But every time I asked I got the same answer: No, I do not know.

At some point, a document materialized with an even longer name than the library: Topkapi Palace Museum Manuscript Library CD Copy Request Form. I now had a form to request a digital copy put on a CD from the library which keeps the manuscripts belonging to the museum of what used to be a palace known as Topkapi. « Notice One » on the form indicated that the form was to be submitted in person, « Notice Four » asked for it to be e-mailed. Of course, Notice One turned out to be correct. And so I deliberated; should I really get on an airplane and spend days just to maybe get digital photos? I thought of Kratchkovsky and Brockelmann, and other scholars of the past, who would not think twice about it, and so I tried my best not to think twice either and plan my trip.

I stayed at the Netherlands Institute in Turkey, on Istiklal Cadessi, in the heart of old Pera. The next day, out I went, to the palace. After I was escorted to the right office inside the palace museum, I was simply asked: “Do you have permission from Ankara?” I had anticipated this question and promptly responded: “I don’t need permission from Ankara for I do not need to see the manuscript itself, I only need photos.” With a big frown, the employee told me that I first needed a permission from the Ministry of Culture and Tourism.

The following day, I went to the office of the ministry which was, of all places, located deep inside a shady, run-down mall, with its entrance next to a tattoo parlor. The guard frowned and waved to the stairs behind him. One floor up, opening a blind door and trying to get the attention of somebody nearby I was again waived up the stairs. On the second floor, I opened a blind door, explained myself as best as I could and this time I was waived toward a blind door down the hallway. I opened it and saw a civil servant watching a television show on his computer, chair reclined, feet on his desk. A colleague of him came, and his German was very good. He asked "Do you have permission from Ankara?" And I responded in the familiar way: "I do not need permission, because I only need photos." He looked through my documentation and told me a third person was needed to sign the permit, who was not there.

The next day I was pleasantly surprised by a phone call: it was signed. Document in hand, I rushed back to Topkapi and was able, this time, to penetrate into the offices of the palace museum library. The remarkably helpful librarian asked me "Do you have permission from Ankara?" I said what I had to say, as I showed all the required documentation. I was after only one manuscript, but had asked for two others, of various texts of Suhrawardī, the Illuminationist philosopher of the 12th century. For these other two manuscripts, I was told, I could come back the next day to receive the CDs. The third one, the one I was actually after, a text by an obscure person named Tūdhī from the 13th century who wrote a commentary on a minor text by Suhrawardī, was to my horror not digitized. "Not to worry," she said while smiling, "we will digitize it and send you the photos by e-mail." Several hundred euros lighter I walked away doubtful they would ever get to it. I would never read this mysterious commentary.

On the plane, Turkish Airlines, I looked out of the window onto Europe. What was I thinking spending so much money on a wild goose chase!? Tūdhī would never show himself to me. History had already gobbled him up. Or could it? Could they actually be digitizing the manuscript right now? As I visualized them busy photographing I became more optimistic. Then, a first rumble rippled through the airplane. Then another. A jerk to the right, like a huge hand had pushed the airplane aside. Waves of turbulence tossed and flicked the airplane in every direction. A free fall of a full two seconds. People screaming, some crying. I crossed myself and despaired: "This manuscript will be my death!"

Philologika Electronica

"If a book is printed, there remains no flavor in it any more," Hoja Ismail once said.[6] This Turkish librarian of the mid-twentieth century is described by Helmut Ritter as "surrounded by flocks of cats, which he loved tenderly. They would be sitting on his lap, his shoulders, his arms, on the heaps of manuscripts

[6] Ritter, H., "Autographs in Turkish Libraries," pp. 63–90 in *Oriens* vol. 6, no. 1 (1953), p. 64.

around him." Hoja Ismail understood the intrinsic value of moving about manuscript folios between your hands. Ritter was effectively exiled from Germany for twenty years, and found new friends in the hundreds, thousands of manuscripts that he met in the libraries of Istanbul. "The struggle between men and books," as Kratchkovsky puts it,[7] was easily won by the books, in the case of Ritter. His series of sixteen articles all entitled *Philologika* are a testament to the riches he found in his newfound friends. Then there is Franz Rosenthal's series of sixteen articles, all entitled *From Arabic Books and Manuscripts*, which is based on the riches he found in Istanbul and beyond. In the years that Henry Corbin was among Istanbul's manuscripts he collected enough materials for about sixteen editions, published throughout the rest of his life. Many, many other scholars passed Istanbul to taste some of that flavor.

I was determined to follow in their footsteps. In 2011, I settled for two months to work day in, day out at the Suleymaniye library, where most manuscripts are now centralized. The director gave me permission to work unto *niṣf al-layl*, which I asked him to repeat two more times because never had I encountered a special collections library open until midnight. A daily routine settled in. Walking down Pera, over the Galata bridge with its fishers, then an immediate right turn, winding up to the Suleymaniye mosque with next to it the library. For lunch going out for a grilled sandwich (*tost*) and then back to work. When my eyes were too tired, I would go back home by the same route, sometimes taking the Tünel train, the second oldest underground rail, to save me from the steep hill up to Istiklal. By night one of the many restaurants and a good conversation or some pages from Pamuk's *Istanbul*.

It was everything I had hoped for, except that my hands did not grace the old paper and vellum of manuscripts, but merely the grubby keys of a keyboard, attached to a computer. In the Suleymaniye library, I spent my working hours looking at a computer screen, figuring out what the best search queries were to get something meaningful out of the digital catalog that was written in a hopelessly unsystematic Turkic transliteration system, while the librarians took another of their many naps. Manuscripts were not to be shown to visitors, only the computer terminals. Despite the general awfulness of that, manuscripts became just one click away to instantly appear, and surely that is a good thing even if it kills Lady Serendipity, who smiled so generously on Corbin one day.[8]

[7] Kratchkovsky, p. 34.

[8] "During the course of a period of work at the Library of Santa Sophia (Aya Sofia), a lucky error in a shelf mark brought me a quite different manuscript from the one I was expecting, but which, in compensation, contained the Persian translation of the Recital of *Hayy ibn Yaqzan* with a commentary in Persian." Corbin, H. *Avicenna and the Visionary Recital*, transl. William R. Trask. New York, United States of America: Pantheon Books, 1960, p. 6.

I focussed on late fifteenth century Ottoman intellectuals, a group of people who remain obscure but whom I had gotten to know from studying with Ihsan Fazlioglu, who has been roaming this library for decades. Dozens and dozens, perhaps hundreds of manuscript whizzed by. Khojazāda's commentary on al-Ghazālī's *Tahāfut*, on which I would publish a few articles, Khojazāda's debate with Mulla Zayrak at the court of Mehmed II, a discussion on whether 'direction' is a real, absolute thing, or only relative, with contributions by Kastalī, Khāṭib Zāde, Sinan Pāshā, Khojazāde, Afḍal Zāde, ʿArab, Qāḍī Zāde, ʿAlāʾ al-Dīn Ṭūsī, Bahāʾ al-Dīn and Nashajī, Khiḍr Bey's didactic poem *al-Qaṣīda al-nūniyya* and the commentary on it by his star-student Khayālī, and on and on.

In my off-hours, I went to places where these people either used to be or where their memory lives on. Machiel Kiel, architectural historian of the region, took me to Mulla Zayrak Cami and the Aya Sophia mosque and kept talking and walking until my ears got tired and my feet hurt. The intellectual epicenter for the folks I was interested in was Fatih Mosque with its eight *madrasas*. Those eight buildings are still there, but closed to visitors and overgrown with weed. Once I convinced a doorman to let me in, we were chased two madrasas down by a ferocious looking guard dog. As the city had evolved for a good five-hundred years, I found I best connected with my new friends through their digitized handwritings.

Only one time did I insist to see and handle a physical manuscript. Literally as they were getting it out, a phone was given over to me with the whisper that it was Professor Fuat Sezgin, the seemingly immortal[9] scholar who had from his younger years right here in this library, built a career that propelled him to head a research institute in Frankfurt, and now a museum in Istanbul. He was at the museum and needed me there this very instant. I obliged, and never did I see the manuscript I was hoping to hold.

The Case of the Missing Word

I was applying final touches to my book *The World of Image in Islamic Philosophy*, when I was reminded of an odd issue in the text of Shahrazūrī's commentary on Suhrawardī's *Ḥikmat al-ishrāq*. Hossein Ziai edited this text while working in Los Angeles in 1993, basing himself on manuscripts from Tehran, Istanbul, and New Haven. In it, a certain sentence reads, literally: "… it is known in what

[9] Lady Serendipity threw an oddball at me while writing this short story. Merely days after I finished it, I got news of the passing of Prof. Sezgin, at age 93. I kept the wording here since it does justice to his personality.

that it is not …". At first, sometime in 2012 while working in Montreal, I translated it as "… it is known that it is not …". I did see the awkwardness of the *fī mā* ("in what") between *qad ʿurifa* and *annahu*, but I could not imagine what else it could mean other than what I translated.

On June 16, 2014, a sunny afternoon in the East of the Netherlands, I shared coffee, cake, and a long discussion on my research with Joep Lameer. He suggested a word was missing in the edition, between *fīmā* and *annahu*, and thought of *taqaddama*. Then the sentence would read "… it is known from what preceded that it is not …". This makes sense, as it is fairly common for a philosopher to refer to a discussion earlier in the book, in which some result was established which is now applied to a different context. "It is known from what preceded" is the typical way to say that. Given the substantial criticism that Lameer gave me that afternoon, the case of the missing word was far from a priority.

Now, on April 7, 2016, in New Haven, Connecticut, it all came back to me as I was revising this very translation, for my supposed book. What to do … what to do … I did not have a digitized manuscript of this text so I felt rather powerless, staring at my computer screen. Then it hit me: that manuscript Ziai used and liked so much, that was in the very same city where I am! I looked it up in the online catalog of Beinecke Library and filed a request. An hour or so later I received news that it was ready. Onwards. Looking around at my stuff I wondered what to take. I squeezed my phone and my wallet in my pockets and off I went.

Sitting down, my experience with manuscripts and my intimate knowledge of the text come together in what seems like an encounter with an old friend. Even though I had not handled this manuscript before, I know it, and the way the folios flip through my hands it seems that the manuscript knows me and is more than happy to let me go my way. I check my phone, on which I have a photo of the page of the edition with the questionable sentence. I check back in the manuscript and flip a little more forward. My eyes glide over the lines—one more folio to turn. There, there it is, and a little moment of triumph: I find the word *marra* scribbled above and in between *fīmā* and *annahu*! *Marra* means 'to pass, come before', which in this context is a perfectly fine synonym for the *taqaddama* that Lameer hypothesized. I turn to the camera function of my phone, snap a picture of the relevant folio, close the manuscript, pet the codex one more time, and give it back to the librarian. I pocket my phone and wallet, and retrace my steps back to my office. My laptop was still open, the cursor still blinking in my word processor. I amend my translation to read "it is known from what « came before » that it is not …". The case of the missing word can be closed.

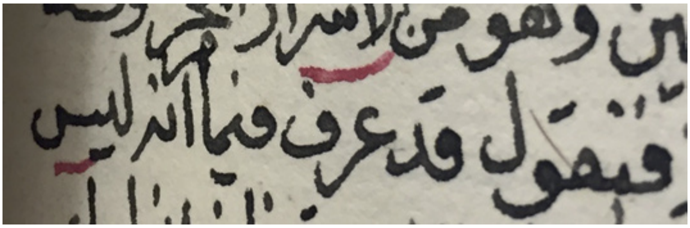

فنقول قد عرف فيما انه ليس

FIGURE 9.1A MS Landberg 7, f. 199b

فنقول: قد عرف فيما انه ليس

FIGURE 9.1B Edition Ziai, p. 509

In the Quiet and Still Air of Delightful Studies

It is past my normal bed time already but I cannot sleep. Outside, Montreal is quietly collecting more snow. I know dozens, hundreds maybe, of men and women are not so quietly at work to remove the snow from the city so that we can go about our normal business tomorrow. A battle between nature and mankind, so to say. I am thinking of my own battle, one with a handful of manuscript copies of a text written in the fifteenth century by an Ottoman philosopher called Khojazāde. The text is supposed to be a commentary on al-Ghazālī's *Tahāfut al-falāsifa* but it is not really a commentary. And neither is it a truly original piece, as I noticed that entire paragraphs coincide with passages from works by earlier, influential intellectuals such as Fakhr al-Dīn Rāzī, Taftāzānī, and Jurjānī. It is not a unique text in another sense, namely, from the exact same time there is another text just like it by another court intellectual called ʿAlāʾ al-Dīn Ṭūsī. The similarity is no coincidence. Sultan Mehmed II, who had just conquered Constantinople, ordered the two to engage in a competition on who could write the better study of Ghazālī's *Tahāfut*. Fast forward to modern times and somehow the book of the loser of the competition, ʿAlāʾ al-Dīn Ṭūsī, had been edited while the winning book by Khojazāde remained unedited. And now I was editing three chapters of Khojazāde's book that together make up the discussion on God's knowledge, especially the question of how God can know particular things if He does not have a body to give him sense perception.

"The game comes running to the hunter," says Kratchkovsky, which had proven to be true in this case. When I looked closer into this debate staged by the sultan, on suggestion of my professor Ihsan Fazlioglu who also gave me some digital photos of relevant manuscripts, editing became an obvious task. But it seems that the game can also lead on the hunter, pulling the hunter deeper into the forest of manuscripts with its labyrinth of marginalia and intertextual connections. On this night, I had stopped working because it was my normal bed time, but the manuscripts were not done with me yet. I knew on a rational level that these manuscripts were in Istanbul, quietly sitting on a shelf in a dark archive. But here they were, in my waking dreams, asking to be lit-up and seen on the screen of my laptop.

It was not like this was a fleeting desire soon to be overcome by sleep. I was in 'The Zone' and I was 'On Fire.' It was only earlier that day that I had cracked the intertextual code of Khojazāde's book and I had also just become more proficient in reading the particular handwriting of the manuscripts. And so, bundled up in my comforter on my couch, I opened my laptop again and plowed through folio after folio, typing out the text while noting differences in the manuscripts and highlighting intertextual relations, until the crack of dawn. The reflection of the moon on the snow and the screen of my laptop were the only sources of light, drawing my attention with laser-sharp focus on the manuscripts and nothing else. A cup of coffee and electronic music pushed my state of mind even further away from reality. "Tonight we are all manuscripts!" I cackled, as my hands raced over the keyboard.

When the snowplow crew came through my street, it must have been between four and five in the morning, I snapped out of my flow. I realized I was looking at the one digitized manuscript that could be Khojazāde's autograph and I wondered: what would Khojazāde think of all of this? What would happen if he just walked through the door and I showed him how I was typing out his book in the middle of the night, based on images of manuscripts, all appearing on this mysterious device, on the other side of the planet, having no real reason to do any of this other than better understanding what was going on five-hundred years ago? My mind shifted to Ghazālī and I wondered, what would it be like to be transported to the eleventh century and meet this great theologian? I would tell him that about five hundred years later two philosophers would write a study on his text and that about five hundred years after that I was studying all these texts based on digital photos. I would show him my iPad and swipe through the manuscripts and my edition. I would look over to Ghazālī and see him look back aghast. Mind. Is. Blown.

A year later I would walk through the madrasa in Istanbul where Khojazāde quite probably wrote his study, and in New Haven I once held a manuscript which is the artifact closest to Ghazālī himself.[10] But during that night in

[10] It seemed even like an autograph, but Frank Griffel spoiled that dream for us. Griffel, F. "Is There an Autograph of Al-Ghazālī in MS Yale, Landberg 318?" pp. 168–186 in *Islam and Rationality*, vol. 2, edited by F. Griffel, Leiden: Brill, 2015.

Montreal, and ever since, I noticed an unbridgeable chasm because as close I was getting to these philosophers of long ago, they were not getting any closer to me. Then what drove me to do what I just did that night? I looked at my laptop and the manuscripts lit up with a beaming smile. Yes, it is not so much Khojazāde and Ghazālī, but these digitized manuscripts which have seduced me, and I let myself be seduced.

A Digital Balm against Schimmelitus

I really lucked out in my undergraduate years. I studied in a time when the Dutch government still gave decent financial support to students, no questions asked as long as you finished your degree. Equally, professors could offer courses without being too bothered by silly requirements like a minimum number of students. And so having started out with more than a hundred fellow students in my first lecture of the beginners course for Arabic, after a few years I was the last one standing. Bernd Radtke, a respectable scholar of Sufism, offered the advanced undergraduate level and since it was just me I was asked to come to his apartment rather than to meet at the university. Sessions would start with a mantra of half-Dutch, half-German complaints about me, the university, the city, society, and life in general. Then Turkish coffee was made. We would sit down at the dining table and get to business. His apartment was filled wall to wall, floor to ceiling, with all the usual reference works and tons of editions, translations, and studies from Islamic studies, with an emphasis on Sufism. We read chapters from Ibn Ṭufayl's *Ḥayy ibn Yaqẓān* and Naṣīr al-Dīn Ṭūsī's *Maṣāriʿ al-muṣāriʿ*. These came from editions, but the text was always held suspect and every letter was overturned, questioned, interrogated. Perhaps he held a slight bias against editions from the Islamic world itself, preferring to use older editions if they came from the hand of a Western scholar. Nonetheless, only a small group of elect scholars were exempted from his contempt for the apparently rampant lack of knowledge of Arabic among editors. "Arabic is a very precise language," Radtke would often times say, as he was figuring out what a *-hu* ('him, his') was referring to and whether it should actually be *-hā* ('her, hers') or whether the verb should be amended. At that time I quietly disagreed with him since it was not clear to me at all why he was saying that a certain word was absolutely the subject or why a sentence definitely was missing a word.[11] I was reading above my pay grade and was already thrilled to get the gist of the sentence. Besides, maybe the original author had made a mistake, that's

[11] Years later, when studying Persian, I appreciated his comment much better.

possible, no? But for Radtke, it was primarily the editor who was at fault. One would think, then, that he would be eagerly looking into manuscripts to check the edition and look for variant readings but he simply was not one to do much with manuscripts. His world was made of paper, print materials.

And so we worked for an hour or two, two times a week, each time making little more progress than half a page. Progress was not only inhibited by the very precise and scholarly manner of reading which he demonstrated to me, but also by the frequent interruptions in which Radtke would get up, scurry off to some shelf somewhere in his house (all the while shouting back at me about the historical or doctrinal context of what we were reading), and come back with an edition or study which he would thump on the table in front of me. He would then say something about the historical figure or the scholar who wrote that book and give me a little moment to browse through it. Invariably, each session a pile would grow on that dining table.

Without being aware of it myself, Radtke was introducing me to the way he as a professional scholar worked. He showed me how his entire life turned around it, basically living not in a house but in a library, and reserving the biggest room of the house not as a living room but for his study. He loudly scorned advertisements seen around the city that promoted a website about 'fun things to do on the weekend' (as though I could do something about it) and I overheard him on the phone with a university administrator, who was nagging him to register his vacation days, fuming that "scholars do not go on vacations!" His love for books and the texts they contain was evident and contagious, and his love-hate relationship with the authors of these books instilled a critical attitude of questioning theirs and my own assumptions and prejudices. Through all of this, admirably, Radtke was selfless enough to not force texts on me from his own expertise in Sufism. Of course he would occasionally mutter how idiotic these philosophers were (and how pathetic my Arabic was) but he encouraged me to make my own choices and never even so much as asked me if I would like to do more with Sufism.

Indeed, one fateful day I shared some possible topics for my undergraduate thesis and mumbled "maybe I want to write something about Ibn ʿArabī, I've read he is a very interesting figure." Radtke turned red in the face and snapped back at me "do you have a case of the Schimmelitus!? Oh you just wait!" And he stormed out, his chair slightly twirling on its feet. By 'Schimmelitus' he was referring to the famous scholar Annemarie Schimmel, whose last name he turned into some kind of a disease playing on the Dutch and German word for mould or mildew. I never pressed him on exactly what that was all about, but it seemed that Schimmel was to him the epitome of an uncritical, romantic approach that does no justice to historical reality. In other words, the exact opposite of what he understood to be sound scholarship. When he came back he was staggering under the weight of four huge tomes that he smacked on the table with a loud bang. It was the Cairo edition of 1911, of Ibn ʿArabī's *al-Futūḥāt al-Makkiyya*, running for about two-thousand pages. "THIS is what you will need to be reading if you want to study Ibn ʿArabī!" Radtke said, visibly annoyed. "This is what you would need to read for the rest of your life! So tell me, do you still want to write about Ibn ʿArabī in your thesis?" "No sir, I don't," I answered.

About eight years later I have got my first book out and I receive the good news that the Dutch government is going to support me for four years, to develop my second book. The topic: Ibn ʿArabī. It is not like I wanted it to be about Ibn ʿArabī—I try to be a man of my word—but my research path simply let me to it. Nor am I planning to read those four bricks of volumes for four years straight. I simply think I don't need to, since my workflow is entirely different from Radtke's. His world is print and mine is digital, and that makes all the difference. At the same time, if I were to show to him what I do, step by step, I do not think he would object. A critical close reading is still at the core of my doings, but it is wrapped in a distant reading method, by which I can whizz between different expressions of the same text (such as a manuscript, an edition, and a translation) and texts of different authors in matter of nanoseconds. For *al-Futūḥāt*, I even have a digitized copy of an autograph. Whereas Radtke puzzles over the correct spelling of a sentence through grammatical and historical knowledge, I switch back and forth between printed editions and manuscripts, I triangulate it by looking at how the text is written and interpreted by ancient commentators, and I check my own findings against modern translations and studies. This difference in workflow has also led to different questions that lead our inquiries. To put it simplistically, Radtke is after the author's intentions whereas I am going for the readers' response. For Radtke it is therefore necessary to stick to one author, preferably one text, and go round and round within that confined space until he can confidently say he has a good grasp of it. For the questions I ask, I need to confine myself to a much smaller unit, for example a chapter or a paragraph, and string together other texts that connect to that, repeating this often enough to be assured that I have sketched out the skeleton of the reception which cannot be drastically changed by future discoveries.

Unbeknownst to me I have integrated a lot of Radtke's workflow. I, too, live in my personal library, which I, too, have filled with books and articles that I love and care for. The difference is that whereas Radtke needed an entire apartment, I have condensed that space to a laptop and some external hard drives. Just as he will browse his shelves, I browse my folders. Just as his living room is his study with his most trusted resources at hand, so my desktop is filled with shortcuts to my research tools. Just as he will collect a pile of books on his table, I will have a pile of PDFs open. And just as I have seen him write notes on index cards, I use a note taking app to organize stray notes. Could it be true that that grumpy man's way of living has rubbed off on me? I really lucked out in my undergraduate years.

Redundancy

"In June of 1835, Baron de Morogue, a member of the Superior Council of Agriculture delivered an address at the French Academy of Sciences [...] 'Every machine,' wrote de Morogue in his report, 'replaces human labor, and therefore every new improvement makes superfluous the work of a certain number of people in industry.'"[12] Now replace *machine* with *digitized manuscript* and *human labor* with *physical manuscript*, and you get what I thought to be true at first. Namely that physical manuscripts are made redundant when they are digitized.

I am reminded of the time I went day in, day out to Beinecke Rare Book & Manuscript Library at Yale University. My aim was to find discussions of cannibalism in Islamic theological texts, if you will believe it, but that is besides the point. Beinecke Library is mostly under the ground, but the entrance is a splendid, very large cubic space whose walls are made of alabaster. When the sun is out, this gives a most extraordinary light inside, which in turn makes visible a smaller cubic shaped construction made of glass and steel. This cube consists of book shelves filled with splendid tomes of rare and valuable books. It is as though the collection is suspended in mid-air, forming a cultural brain. But if the books are the brain, then I as a visitor am only a passing thought. I digress. What I mean to say is that one day, all those shelves were completely empty. For a regular visitor this was a startling sight. What happened? I smiled at the thought that all books must have been digitized and upper management, seeing that the books themselves were now superfluous, redundant, ordered them to be burned. Soon all those shelves would be converted to server space and there would be a buzzing brain of super computers suspended inside Beinecke's cube.

I have come to realize this would be madness. One extraordinary thing about manuscripts is their ruggedness. If a manuscript of eight-hundred years

[12] As cited in Roerich, N. *The Invincible*. New York: Nicholas Roerich Museum, 1994, p. 119.

old is still in its original binding, it will easily survive another eight-hundred years. Well, I do not know that for sure, but that is the impression they give me. I have seen manuscripts for which I would not be surprised that were I to drive over them with a tank, the tank would be damaged, not the manuscript. So, as long as you throw them in a v-shape on some pillows, you can be pretty rough with them. You can flip the folia through your fingers for a great browsing experience, or use a magnifying glass to painlessly zoom in on something. Above else, they are entirely self-contained; to use them the only thing you need is the manuscript itself. Compare that with digital files! Alter one character in the encoding of the file and the entire thing becomes corrupted. And with a push of a button the file is simply deleted and irrecoverable. Browsing is usually clunky, as you have to swipe up and down and usually completely miss the intended page. Zooming in will only go so far as the resolution of the photo allows you to. Above all else, digitized manuscripts need a hardware medium to subsist, and a whole set of hardware and software components to be used. Magnetic tape (still used by businesses for storing backups), if left alone on a shelf, will hold its data at most for thirty years. CDs and DVDs, if kept under laboratory-like ideal circumstances, will hold their data for at most twenty years. Flash storage, like SD cards, will last for at most ten years. Hard drives, finally, like the one in your computer, will typically not make it past five years before failing. Digital persistence is a joke. So whereas physical manuscripts can be shelved somewhere, forgotten for centuries, and then still be used, digitized manuscripts require continuous attention and rejuvenation.

It has happened plenty of times now, that I acquired another big batch of them, for example given to me on an external hard disk. I connect the hard disk for the first time to my laptop and I find the manuscripts organized in folders with collection names, then each manuscript is merely a number ending in the familiar .PDF extension. I scroll through them as an army officer marching up and down the square inspecting the new recruits. I stop randomly at several ones inspecting them more closely, to see their file size, if they can be opened, and to get an impression of the image quality. As I scroll through the list, the hard disk spins up and buzzes. I can hear the manuscripts whisper: 'sustain me and I shall be saved, and ever observe your commands.'[13]

After inspection, I welcome them into my family and commit to care for them. They are filed in the right directory with the right name. If they are especially important they are brought into a multi-layered system of offline and online backups. And even if they are not, they are copied onto another medium, for redundancy. That's right, in the digital world, for something to be

[13] Psalms, 118:117.

sustained, it needs to become redundant. Two copies of the exact same file are stored on different devices, so that if one of them goes corrupt or the hardware on which it is stored goes bust, the other copy is still there and can be copied to a new device to again comply with redundancy. This is not forgery nor is there anything inauthentic about it, since every instance of the same file is equally the original as well as the copy. Each are redundant, vincible, under constant threat of going out of existence. Only by the frequent handling of a reader can they survive. Physical manuscripts, on the other hand, are invincible. It is exactly frequent handling that can threaten that. I suppose that this is why manuscripts remain at a distance and only sometimes call out to readers to come over and learn of their secrets. Digitized manuscripts, on the other hand, come running towards the reader. You cannot 'have' digitized manuscripts as you can have physical manuscripts lying around on a shelf or in a box. In order to have digitized manuscripts, you need to be among digitized manuscripts.

Leiden or 578

A manuscript came onto my radar, a copy of an unedited commentary by Shahrazūrī (d. ≥ 1288) on Suhrawardī's (d. 1191) *al-Talwīḥāt*. It was kept at the library of Leiden University, famous for its Islamic manuscripts. On a summer day, I went to Leiden to look at all manuscripts relevant to my project, and snapped some photos of this manuscript of interest, *Leiden Or 578*. I was lucky enough to break up my day among Leiden's manuscripts with a lunch with Professor Jan-Just Witkam. He was very generous with his knowledge, a similar generosity he shows through his website www.islamicmanuscripts.info. I had previously studied with Adam Gacek, so after this lunch I only needed to meet François Déroche to complete the holy trinity of Islamic Manuscript Studies.

For about a year, I made good and pleasant use of my digital photos of *Leiden Or 578*. However, the time had come to move elsewhere, much further away from Leiden than a simple train ride, and the manuscript was still on my mind. So I returned to *Leiden Or 578*, this time armed with a quality DSLR camera, intend on digitizing the entire manuscript.

I collect my manuscript and walk over to the reading table. I pass someone who is carefully measuring the length of a manicule, if I remember correctly, in a margin of what looks like a veritably ancient manuscript. It is a sweet sight, to see reader and manuscript working closely together. Sitting down at the reading table I first take my time to get reacquainted with my buddy. It is such a big, beautiful manuscript, with 270 large and thick pages, in a sturdy binding. Built like a tank. Something that is going on next to me, on my right, is distracting me. Two people are talking louder than needed—don't they know that manuscripts require silence? I soon realize where they get the audacity from; one of them works at the library and therefore thinks she owns the place. This is all fine and well, but they are placing and pushing around handwritten, single pages directly on the table, and they mix up the stack of loose pages that they got. An urge builds up to say something about it, to point out the irreplaceable nature of the artifacts they hold and to remind them of the etiquette of leaving an artifact behind in the order that you found it. The quiet one packs up and gets ready to leave before my urge reaches a boiling point. Meanwhile, the librarian, the loud-talker, notices me. "Whats that!?" the loud-talker says, holding up a real-life manicule to my camera. Before I open my mouth she continues: "you can only take twenty photos with that!" I am a bit baffled. Next to the place, loud-talker owns my future photos as well? I have not heard of this rule before. Why would the library care that I make photos for personal use, of a couple of 700-year old pieces of paper on which is scribbled some dull philosophy? Would they stop me from copying the manuscript by hand, beyond twenty pages? I let out a sound as uncommitted and neutral as I can and fall back in my chair. In the corner of my eyes, I keep watch of the front desk.

After a while, loud-talker noisily announces she will take her coffee break and steps out. 'Perfect,' I think to myself. I pick up my camera and start snapping. The metadata of my images tell me that the photo of the front cover was taken at 9.35 AM. The last shot, of the back of the codex, is taken at 10.35 AM. In one hour, I took almost three-hundred photos, or around five per minute, at every step taking care to not harm my dear *Leiden Or 578*. Now that I have the manuscript digitized, you know what that means. Time to go. I give back my buddy to the front desk and make my way through the corridors. Near the exit of the library, I see loud-talker coming back from her hourlong coffee break. I nod and smile as we pass.

Dark Archive Fever

It is late in the evening and I have stumbled upon a very large collection of manuscripts on the internet. "The wealth of manuscripts overwhelmed and fascinated me," as Kratchkovsky would say.[14] I set out to download them, using right-click and selecting *Save*. I think it is better to have them, stored away

[14] Kratchkovsky, p. 37.

locally on a hard drive, than to not have them. Ibn Taymiyya writes that “those who were present at the deathbed of Khūnajī, the chief logician of his time, reported that just before his death he said: ‘I die knowing nothing except that the possible presupposes the necessary.’ He then added: ‘And presupposition is a negative attribute, so I die knowing nothing.’”[15] There is this trend to store everything in ‘the cloud’ which is a fancy word for handing over the data you own to big corporations. And ‘handing over’ is a negative attribute so I end up owning nothing.

The first attempt fails. I think it has to do with the file size, which make the server time out. I reach for a certain piece of software and give it instructions to download the manuscripts. I think it is simply better to have them on my own computer. Walter Benjamin has said that “the presence of the original is the prerequisite to the concept of authenticity.”[16] When I read a translation I look for editions to present themselves, and when I read an edition I look for manuscripts to present themselves. When I studied at McGill, I had a designated carrel in the library. I put all kinds of book on there, and one day I was reading a letter by the famous theologian Fakhr al-Dīn Rāzī to a friend, and in it he referred to a passage of a certain, lesser-known work by Ibn Sīnā. I looked up and saw exactly that book sitting on the shelf of my carrel, literally within hand’s reach. I suddenly felt so close to Rāzī, like he was talking directly to me. I want it to be like that always. Immersing myself among digitized manuscripts, I think I can get very close to it.

As for this download tonight, it has once more failed. The files are too big or the server is too unreliable. Time to pull out the big gun; a command line tool called *wget*. I do not like to use it because a command line tool only works from the Terminal (Command Prompt on Windows), meaning that you have to type out a command in this black and white screen, with no graphical interface and no mouse functionality. I end up with an incomprehensible command that looks like *wget -r -H -nc -np -nH --cut-dirs=1 -A .pdf -e robots=off -l 1 https://URLofManuscripts* and there is little else I can do than hit Enter and hope for the best. I really do want to have those manuscripts. Plotinus said in the Arabic version of the Enneads: “Asked about the gain (he had) from his love of knowledge, he said: In sorrow, it is my solace; in comfort, my pleasure; in times

[15] Ibn Taymīya. *Ibn Taymiyya against the Greek Logicians*. Translated by W.B. Hallaq. Oxford: Clarendon Press, 1993, p. 132–133.

[16] Benjamin, W. “The Work of Art in the Age of Mechanical Reproduction.” pp. 217–51 in *Illuminations: Essays and Reflections*, translated by H. Zohn. New York: Schocken Books, 1969, p. 220.

of inertia, my stimulus; in times of activity, my tool; in dark hours my light, and when they are gone, my recreation and joy."[17] So it is with my digitized manuscripts. Other people may be washing their car and waxing it, and enjoy their time doing it; I care for my manuscripts by downloading, sorting, and copying. Digitized manuscripts are in a sense quite demanding and it is odd, then, that individual attention is exactly something that can hardly be given to them. For, while physical manuscripts hardly ever congregate in more than four digit numbers, and then only in the great libraries of this world, collecting digitized manuscripts (without a specific purpose) only becomes interesting when they come in five digit numbers or more. Once downloaded, they are stored away in a growing archive whose contents remains largely dark to me. Even when I look for and find a manuscript, I can lose track of it if I am not careful. One time I found the name of Suhrawardī, the philosopher I was investigating at the time, beautifully written together with his epithet 'al-Maqtūl'. I cut it out and saved it, but later when I wanted to use it in promotional materials and I needed to cite its origins, I could no longer trace it back to its origin. It retreated back into the darkness of the digital archive. This was not only a shame but also terribly annoying, because there was a scholarly argument to be made here. Some scholars have insisted that Suhrarwardi was referred to by admirers as "al-shahīd", meaning 'the martyred one,' but I think his followers used the same epithet everybody used, namely "al-maqtūl", meaning 'the killed one', given to him since he was put to death by the local ruler of his time. This beautifully rendering is proof for it, but if I cannot cite it, it is like it is not there.

My computer is churning away. I want to keep an eye on it but I also want to sleep. So I pull it up next to my bed and turn the brightness of the screen down to zero. This *wget*-method of downloading seems to be slow but steady. I am not entirely sure but I think these manuscripts I am steadily pulling in are stored on a server in no less than Mecca. This reminds me of Nicholas Roerich words: "Once when I was asked, 'What is the difference between East and West?' I said, 'The best roses of East and West are equally fragrant.'"[18] So it is for digitized manuscripts. It is a thought that lingers on my mind, until the steady crackling of my computer rocks me to sleep.

[17] Plotinus. *Plotini Opera. Enneades IV–V* [*Plotiniana Arabica Ad Codicum Fidem Anglice Vertit*]. Edited by P. Henry and H.R. Schwyzer. Translated by G. Lewis. Paris: Desclée de Brouwer et Cie, 1959, p. 478.

[18] Roerich, p. 126.

Gratitude

Being grateful means you show your appreciation for the kindness of others. Often times, you do this because there is nothing else you can do or give. There have been times that I needed a specific manuscript, from Istanbul, Damascus, Tehran, or Lahore. These are places that are hard or inadvisable to go to as a foreigner. It cannot hurt, of course, to send an email to the library, but often times I was left stranded in bureaucratic red tape.

I recall a particularly long episode with a library in Lahore. After e-mailing a few different people at the library, I got a response. This first point of contact was around the same time when I got my first smartphone. I remember being absolutely amazed when my phone bleeped and a message all the way from Pakistan appeared on my screen. After one and a half years of e-mailing and snail mailing back and forth, I was impressed at what could be done from a distance. However, it still took me a local friend to physically go there and make the final arrangements. As much as I meticulously planned that last phase and tried my best to involve my friend as little as possible, he still ended up paying the fee out of pocket and of course refused to be reimbursed no matter what I said or did. Gratitude can sometimes be very frustrating. Nonetheless, it was a great moment of triumph when the PDF of that digitized manuscript dropped in my email inbox.

Friends and colleagues in Iran are oftentimes much more industrious. For as much as Iranian culture 'stands on ceremony,' they have no qualms including a finder's fee in the total amount due if you ask how you can compensate them. I on my part take this as a cue to not bother with the libraries myself but simply give these people the full details of the manuscript, and let them do the rest. One to three weeks will go by and sooner or later a PDF will appear in my inbox. Paying for this is pretty much exactly what money is for, and yet it does make gratitude a bit awkward at times.

Then there is war-ravaged Syria, where I did not even think it was appropriate to start asking around people for favors. And so I kept my wishes silently with me for years. There was a time when ISIS burned and sold manuscripts and other cultural heritage objects. I was sometimes asked "is it not awful what they are doing there? The catastrophe! These artifacts are priceless!" Priceless they may be, but every second they waste their time with it is a second not wasted on torturing people. And actual people are far more priceless than these manuscripts, these manuscripts which we by and large did not look at anyway so why all of a sudden should we wish to save them? Unexpectedly, a colleague gave me a batch of his files among which I found a digital copy of the manuscript from Damascus I had in mind. Who was I to thank now? Sure

I thanked the colleague, but not for this find but only for the batch in total. He had, after all, no clue this one file was an important piece for me to complete a puzzle I was working on. Meanwhile, the people in the Zahiriyah library who had digitized the manuscript (or, rather, the microfilm of the manuscript), remained unknown and anonymous to me. So here we are. I got my stupid manuscript like a spoiled child who got his way, and I had nothing to offer in return as I could not even thank the relevant people. Gratitude can leave you sorely dissatisfied.

For Istanbul, too, it has at times been better to use an intermediary. Manuscripts from Turkey are popular, not only because many of them are excellent and precious, but equally so because they have been mass-digitized and mass-catalogued in a usable manner. That is not to say this has been *usefully* done, just that the way it is done is *usable*. Gratitude means you are happy with what you get and do not complain about what it is not.

Once in a while somebody will ask me who my contacts are so that they can request a manuscript, but I never give it out. These people in Turkey, Syria, Iran, and Pakistan do it as a service to a friend, even if they get money for it, and it seems ungrateful to send others their way as though this is a public service they offer. With every request, it remains a delicate balance who and when to ask. Nevertheless, the words of Kratchkovsky ring true today as they did a hundred years ago: whenever I come across an interesting or enigmatic comment about a manuscript, “my curiosity was aroused and I sought to obtain a photograph of the manuscript,”[19] wherever it may be on this planet.

[19] Kratchkovsky, p. 136.

Bibliography

Abdulrazak, F.A. “The Kingdom of the Book: The History of Printing as an Agency of Change in Morocco between 1865 and 1912.” PhD dissertation, Boston University, 1990.

Adams, T.R., and N. Barker. “A New Model for the Study of the Book.” pp. 5–44 in *A Potencie of Life: Books in Society*, edited by N. Barker. London: British Library, 1993.

Ahmed, S. *What Is Islam? The Importance of Being Islamic*. Princeton: Princeton University Press, 2015.

Ainsworth, P. “E-Science for Medievalists: Options, Challenges, Solutions and Opportunities.” *Digital Humanities Quarterly* 3, no. 4 (2009).

Akkerman, O. “The Bohra Dark Archive and the Language of Secrecy: A Codicological Ethnography of the Royal ʿAlawī Bohra Library in Baroda.” PhD dissertation Freie Universität. Berlin, 2015.

Albin, M.W. “Printing of the Qurʾān.” vol. 4, pp. 264b–276b in *Encyclopædia of the Qurʾān*, edited by J.D. McAuliffe, 6 vols., Leiden: Brill, 2004.

Althusser, L. “Ideology and Ideological State Apparatuses.” pp. 127–88 in *Lenin and Philosophy and Other Essays*, translated by B. Brewster. New York: Monthly Review Press, 1971.

Anderson, B. *Imagined Communities*. 2nd ed. London: Verso, 2006 [1983].

Andrews, T. “The Third Way: Philology and Critical Edition in the Digital Age.” pp. 61–76 in *Variants* 10 (2013).

Andrews, T. “Digital Techniques for Critical Edition.” pp. 175–195 in *Armenian Philology in the Modern Era: From Manuscript to Digital Text*, edited by V. Calzolari and M.E. Stone. Leiden: Brill, 2014.

Andrews, T., and C. Macé, eds. *Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches*. Turnhout: Brepols, 2014.

Apollon, D., C. Bélisle, and P. Régnier, eds. *Digital Critical Editions*, Urbana: University of Illinois Press, 2014.

Arnold, D. “Digital Artefacts: Possibilities and Purpose.” pp. 159–70 in *The Virtual Representation of the Past*, edited by M. Greengrass and L. Hughes. Farnham: Ashgate, 2008.

Arsene, C.T.C., P.E. Pormann, N. Afif, S. Church, and M. Dickinson. “High Performance Software in Multidimensional Reduction Methods for Image Processing with Application to Ancient Manuscripts.” pp. 1–25 in *Manuscript Cultures*, 2016.

Atiyeh, G.N., ed. *The Book in the Islamic World: The written word and communication in the Middle East*. Albany: SUNY Press, 1995.

Aussems, M., and A. Brink. “Digital Palaeography.” pp. 293–308 in *Kodikologie und Paläographie im digitalen Zeitalter*, edited by M. Rehbein, P. Sahle, and T. Schaßan. Norderstedt: BoD, 2009.

Babeu, A. “Rome Wasn’t Digitized in a Day”: Building a Cyberinfrastructure for Digital Classics. Washington: Council on Library and Information Resources, 2011.

Baker, C. “Editing Medieval Texts.” vol. 1, pp. 427–450 in *Handbook of Medieval Studies: Terms—Methods—Trends*, edited by A. Classen, 3 vols. Berlin: De Gruyter, 2010.

Barkeshli, M. “Material Technology and Science in Manuscripts of Persian Mystical Literature.” pp. 187–214 in *Manuscript Cultures* 8 (2015).

Barthes, R. “The Death of the Author.” *Aspen* 5–6 (1967).

Baudrillard, J. *Simulations*. Translated by P. Foss, P. Patton, and P. Beitchman. Semiotext[e], 1983.

Bausi, A. (et al), ed. *Comparative Oriental Manuscript Studies: An Introduction*. Hamburg: COMSt, 2015.

Beal, P. *A Dictionary of English Manuscript Terminology 1450–2000*. Oxford: Oxford University Press, 2008.

Benjamin, W. “The Work of Art in the Age of Mechanical Reproduction.” pp. 217–251 in *Illuminations: Essays and Reflections*, translated by H. Zohn. New York: Schocken Books, 1969.

Berger, P.L., and T. Luckmann. *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. London: Penguin Books, 1966 [Reprint 1991].

Besek, J.M., et al., “Digital Preservation and Copyright: An International Study,” pp. 104–111 in *The International Journal of Digital Curation*, vol. 2, no. 3 (2008).

Bischoff, B. Paläographie Des Römischen Altertums Und Des Abendländischen Mittelalters. Berlin: Schmidt, 1979.

Blair, A. “Afterword: Rethinking Western Printing with Chinese Comparisons.” pp. 349–361 in *Knowledge and Text Production in an Age of Print: China, 900–1400*, edited by L. Chia and H. De Weerdt, Leiden: Brill, 2011.

Blatty, W.P. *The Exorcist: A Novel*. New York: HarperCollins, 2011 [1971].

Bobzin, H. “Von Venedig Nach Kairo: Zur Geschichte Arabischer Korandrucke.” pp. 151–76 in *Sprachen Des Nahen Ostens Und Die Druckrevolution. Eine Interkulturelle Begegnung*, edited by G. Roper, D. Glass, and E. Hanebütt-Benz. Westhofen: WVA Verlag Skulima, 2002.

Bodard, G. “EpiDoc: Epigraphic Documents in XML for Publication and Interchange.” pp. 101–118 in *Latin on Stone: Epigraphic Research and Electronic Archives*, edited by F. Feraudi-Gruénais. Lanham: Lexington Books, 2010.

Bolter, J.D. *Writing Space: Computers, Hypertext, and the Remediation of Print*. Mahwah: Lawrence Erlbaum, 2001.

Bond, S. “Dear Scholars, Delete Your Account At Academia.Edu.” *Forbes*, January 23, 2017.

"Books Under the Microscope." *UT News: The University of Texas at Austin*, October 18, 2012.

Bourgain, P. "The Circulation of Texts in Manuscript Culture." pp. 140–159 in *The Medieval Manuscript Book: Cultural Approaches*, edited by M. Van Dussen and M. Johnston, Cambridge: Cambridge University Press, 2015.

Bozzi, A., and S. Calabretto. "The Digital Library and Computational Philology: The BAMBI Project." pp. 269–85 in *Research and Advanced Technology for Digital Libraries*, edited by C. Peters and C. Thanos. Berlin: Springer, 1997.

Bradley, J. "No Job for Techies: Technical Contributions to Research in the Digital Humanities." pp. 11–25 in *Collaborative Research in the Digital Humanities*, edited by M. Deegan and W. McCarty. London: Routledge, 2012.

Brey, A., and E. Muhanna. "Quantifying the Quran." pp. 151–173 in *The Digital Humanities and Islamic & Middle East Studies*. Berlin: De Gruyter, 2016.

Burdick, A., J. Drucker, P. Lunenfeld, T. Presner, and J. Schnapp. *Digital_humanities*. Cambridge Mass.: The MIT Press, 2012.

Busch, H., and S. Chandna. "ECodicology: The Computer and the Mediaeval Library." pp. 3–23 in *Kodikologie Und Paläographie Im Digitalen Zeitalter 4*, edited by H. Busch, F. Fischer, and P. Sahle. Norderstedt: BoD, 2017.

Calabretto, S., and A. Bozzi. "The Philological Workstation BAMBI (Better Access to Manuscripts and Browsing of Images)." *Journal of Digital Information* 1, no. 3 (1998).

Cameron, F. "Beyond the Cult of the Replicant: Museums and Historical Digital Objects—Traditional Concerns, New Discourses." pp. 49–75 in *Theorizing Digital Cultural Heritage: A Critical Discourse*, edited by F. Cameron and S. Kenderdine. Cambridge Mass.: The MIT Press, 2007.

Carroll, L. *Through the Looking-Glass, and What Alice Found There*. London: MacMillan, 1872.

Cartelli, A., and M. Palma. "Digistylus—An Online Information System for Palaeography Teaching and Research." pp. 123–134 in *Kodikologie und Paläographie im digitalen Zeitalter*, edited by M. Rehbein, P. Sahle, and T. Schaßan. Norderstedt: BoD, 2009.

Cartelli, A. "DigiStylus: A Socio-Technical Approach to Teaching and Research in Paleography." pp. 741–753 in *Issues in Informing Science and Information Technology* 6 (2009).

Carter, J., and N. Barker. *ABC for Book Collectors*. 8th ed. London: Oak Knoll Press, 2004 [Or. 1952].

Cerquiglini, B. *Éloge de la variante*. Paris: Éd. du Seuil, 1989.

Chevallier, P., L. Rioust, and L. Bouvier-Ajam. "Consultation of Manuscripts Online: A Qualitative Study of Three Potential User Categories." *Digital Medievalist* 8 (2013).

Chia, L., and H. De Weerdt. "Introduction." pp. 1–32 in *Knowledge and Text Production in an Age of Print: China, 900–1400*, edited by L. Chia and H. De Weerdt. Leiden: Brill, 2011.

Christlein, V., M. Gropp, and A. Maier. "Automatic Dating of Historical Documents." pp. 151–164 in *Kodikologie und Paläographie im digitalen Zeitalter 4*, edited by H. Busch, F. Fischer, and P. Sahle. Norderstedt: BoD, 2017.

Chu, Y., D. Bainbridge, M. Jones, and I. Witten. "Realistic Books: A Bizarre Homage to an Obsolete Medium?" pp. 78–86 in *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, New York: ACM, 2004.

Collins, E., M.E. Bulger, and E.T. Meyer. "Discipline Matters: Technology Use in the Humanities." pp. 76–92 in *Arts & Humanities in Higher Education* 11, no. 1–2 (2011).

Collins, E., and M. Jubb. "How Do Researchers in the Humanities Use Information Resources?" pp. 176–87 in *Liber Quarterly* 21, no. 2 (2012).

Conway, M.E. "How Do Committees Invent?" pp. 28–31 in *Datamation* 14, no. 5 (1968).

Corbin, H. *Avicenna and the Visionary Recital*, transl. William R. Trask. New York, United States of America: Pantheon Books, 1960.

Correa, A.C. "Palaeography, Computer-Aided Palaeography and Digital Palaeography: Digital Tools Applied to the Study of Visigothic Script." pp. 247–72 in *Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches*, edited by T. Andrews and C. Macé. Turnhout: Brepols, 2014.

Correa, D.J. "Digitization: Does It Always Improve Access to Rare Books and Special Collections?" pp. 177–79 in *Digital Technology & Culture* 45, no. 4 (2017).

Craig-McFeely, J., and A. Lock. *Digital Restoration Workbook*. Oxford: OSSC Publications, 2006.

Craig-McFeely, J. "Finding What You Need, and Knowing What You Can Find: Digital Tools for Palaeographers in Musicology and Beyond." pp. 307–39 in *Kodikologie Und Paläographie Im Digitalen Zeitalter 2*, edited by F. Fischer, Chr. Fritze, and G. Vogeler. Norderstedt: BoD, 2010.

Crane, G., A. Babeu, D. Bamman, L. Cerrato, and R. Singhal. "Tools for Thinking: ePhilology and Cyberinfrastructure." pp. 16–26 in *Working Together or Apart: Promoting the Next Generation of Digital Scholarship*. Washington, D.C.: Council on Library and Information Resources, 2009.

Dagenais, J. *The Ethics of Reading in Manuscript Culture: Glossing the Libro de Buen Amor*. Princeton: Princeton University Press, 1994.

Dahlström, M. "Critical Editing and Critical Digitisation." pp. 79–98 in *Text Comparison and Digital Creativity*, edited by W. van Peursen, E.D. Thoutenhoofd, and A. van der Weel. Leiden: Brill, 2010.

Darnton, R. "What Is the History of Books?" pp. 65–83 in *Daedalus* 111, no. 3 (1982).

Deckers, D., and C. Glaser. "Zum Einsatz von Synchrotronstrahlung Bei Der Wiedergewinnung Gelöschter Texte in Palimpsesten Mittels Röntgenfluoreszenz." pp. 181–90 in *Kodikologie Und Paläographie Im Digitalen Zeitalter 2*. Norderstedt: Books on Demand, 2010.

Dekker, R.H., D. van Hulle, G. Middell, V. Neyt, and J.J. van Zundert. "Computer-Supported Collation of Modern Manuscripts: CollateX and the Beckett Digital Manuscript Project." pp. 452–470 in *Literary and Linguistic Computing* 30, no. 3 (2015).

Derrida, J. *Archive Fever: A Freudian Impression*. Translated by E. Prenowitz. Chicago: The University of Chicago Press, 1996.

Dhawqi, F. "Ibrāhīm Mutafarriqa, Risāle wasīle-ye al-ṭibāʿa wa-tarjama ān." pp. 234–282 in *Payām-i Bihāristān* 2, no. 4 (2016).

Dijk, A. van. "Early Printed Qur'ans: The dissemination of the Qur'an in the West." pp. 136–143 in *Journal of Qur'anic Studies* 7, no. 2 (2005).

Dormolen, H. van. *Richtlijnen Preservation Imaging Metamorfoze*. Den Haag: Koninklijke Bibliotheek, 2012.

Driscoll, H. "The Legendary Legacy: Crunching 600 Years of Saga Manuscript Data." pp. 71–79 in *Kodikologie Und Paläographie Im Digitalen Zeitalter 4*, edited by H. Busch, F. Fischer, and P. Sahle. Norderstedt: BoD, 2017.

Dunleavy, P. *Authoring a PhD: How to Plan, Draft, Write and Finish a Doctoral Thesis or Dissertation*. New York: Palgrave Macmillan, 2003.

Easton Jr., R.L., and W. Noël. "The Multispectral Imaging of the Archimedes Palimpsest." pp. 39–49 in *Gazette Du Livre Médiéval* 45 (2004).

Eckstein, L.N. "Of Scribes and Scripts: Citizen Science and the Cairo Geniza." pp. 208–214 in *Manuscript Studies* 3, no. 1 (2018).

Eggert, P. "The Book, the E-Text and the 'Work-Site.'" pp. 63–82 in *Text Editing, Print and the Digital World*, edited by M. Deegan and K. Sutherland. Surrey: Ashgate, 2009.

Eisenstein, E. *The Printing Press as an Agent of Change: Communications and Cultural Transformations in Early-Modern Europe*. 2 vols. Cambridge: Cambridge University Press, 1979.

Eliot, S., and J. Rose. *A Companion to the History of the Book*. Oxford: Blackwell Publishing, 2007.

Emerson, L. *Reading Writing Interfaces: From the Digital to the Bookbound*. Minneapolis: University of Minnesota Press, 2014.

Erickson, H.M., and J. Ogburn. "RetroReveal.Org: Semi-Automated Open-Source Algorithms and Crowdsourcing Tools for the Discovery, Characterization and Recovery of Lost or Obscured Content." p. 80 in *Archiving 2012*. Copenhagen: Society for Imaging Science & Technology, 2012.

Evans, H. *Do I Make Myself Clear?* London: Little, Brown, 2017.

Favilla, E.J. *A World Without "Whom."* London: Bloomsbury, 2017.

Finkelstein, D., and A. McCleery. *An Introduction to Book History*. London: Routledge, 2005.

Fisher, M. "Authority, Interoperability, and Digital Medieval Scholarship." pp. 955–964 in *Literature Compass* 9, no. 12 (2012).

Flügel, G. *Corani Textus Arabicus*. Leipzig: Sumtibus Ernesti Bredtii, 1869.

Foucault, M. *The Order of Things: An Archeology of the Human Sciences*. New York: Vintage Books, 1994 [Or. 1966].

Foucault, M. *The Archeology of Knowledge*. Translated by A.M.S. Smith. New York: Pantheon Books, 1972 [Or. 1969].

Foys, M.K. "Medieval Manuscripts: Media Archaeology and the Digital Incunable." pp. 119–39 in *The Medieval Manuscript Book: Cultural Approaches*, edited by M. Van Dussen and M. Johnston, Cambridge: Cambridge University Press, 2015.

Goodman, N. *Ways of Worldmaking*. Indianapolis: Hackett Publishing, 1978.

Gottfried, B., M. Wegner, M. Spano, and M. Lawo. "Abbreviations in Medieval Latin Handwriting." pp. 3–9 in *Manuscript Cultures* 7 (2013).

Gratien, C., M. Polczyński, and N. Shafir. "Digital Frontiers of Ottoman Studies." pp. 37–51 in *Journal of the Ottoman and Turkish Studies Association* 1, no. 1–2 (2014).

Griffel, F. "Is There an Autograph of Al-Ghazālī in MS Yale, Landberg 318?" pp. 168–186 in *Islam and Rationality*, vol. 2, edited by F. Griffel, Leiden: Brill, 2015.

Griffitts, T.A. "Software for the Collaborative Editing of the Greek New Testament." PhD dissertation, University of Birmingham, 2017.

Gumbrecht, H.U. *Production of Presence: What Meaning Cannot Convey*. Stanford: Stanford University Press, 2003.

Haltrich, M., E. Kapeller, and J.A. Schön, eds. *From Sheep to Shelf: An Illustrated Guide to Medieval Manuscripts for Students*. DEMM, 2017.

Halwaji, A.S. (ed.), *Fihris al-makhṭūṭāt al-ʿarabīya bi-dār al-kutub al-miṣrīya: al-majāmīʿ*, London: Muʾassasat al-furqān li-l-turāth al-islāmī, 2011.

Hassner, T., M. Rehbein, P. Stokes, and L. Wolf, eds. "Computation and Palaeography: Potential and Limits." pp. 1–30 in *Kodikologie Und Paläographie Im Digitalen Zeitalter 3*. Norderstedt: BoD, 2015.

Hayles, N.K. *Writing Machines*. Cambridge Mass.: The MIT Press, 2002.

Hendrickson, J., Adil, S., "A Guide to Arabic Manuscript Libraries in Morocco: Further Developments," pp. 1–19 in *MELA Notes* 86 (2013).

Herzog, R., A. Solth, and B. Neumann. "Computer-Based Stroke Extraction in Historical Manuscripts." pp. 14–24 in *Manuscript Cultures* 3 (2010).

Herzog, R., A. Solth, and B. Neumann. "Computer Methods for Comparing the Hands of Manuscripts." pp. 169–175 in *Manuscript Cultures* 4 (2011).

Hirschkind, C. "Media and the Qurʾān." vol. 3, pp. 341b–349b in *Encyclopædia of the Qurʾān*, edited by J.D. McAuliffe, 6 vols., Leiden: Brill, 2004.

Hirtle, P.B. "The Impact of Digitization on Special Collections in Libraries." pp. 42–52 in *Libraries & Culture* 37, no. 1 (2002).

Hollaus, F., M. Gau, R. Sablatnig, W.A. Christens-Barry, and H. Miklas. "Readability Enhancement and Palimpsest Decipherment of Historical Manuscripts." pp. 31–46 in *Kodikologie Und Paläographie Im Digitalen Zeitalter* 3. Norderstedt: Books on Demand, 2015.

Hörnschemeyer, J. "Textgenetische Prozesse in Digitalen Editionen." PhD dissertation, Universität zu Köln, 2013.

Hunt, L., M. Lundberg, and B. Zuckerman. "Concrete Abstractions: Ancient Texts as Artifacts and the Future of Their Documentation and Distribution in Their Digital Age." pp. 149–172 in *Text Comparison and Digital Creativity*, edited by W. van Peursen, E.D. Thoutenhoofd, and A. van der Weel. Leiden: Brill, 2010.

Ibn Taymīya. *Ibn Taymiyya against the Greek Logicians*. Translated by W.B. Hallaq. Oxford: Clarendon Press, 1993.

Ingold, T. *Lines: A Brief History*. London: Routledge, 2007.

Jannidis, F., H. Kohle, and M. Rehbein. *Digital Humanities: Eine Einführung*. Stuttgart: J.B. Metzler, 2017.

Jawharī. *Tarjama ṣiḥāḥ al-Jawharī*. Translated by Vānqulī, printed by Ibrāhīm Müteferrika. 2 vols., Istanbul: Dār al-ṭibaʿa, 1141h (1729).

Jeanneney, J.-N. *Google and the Myth of Universal Knowledge: A View from Europe*. Translated by T.L. Fagan. Chicago: The University of Chicago Press, 2007.

Johnston, M., and M. Van Dussen. "Introduction: Manuscripts and Cultural History." pp. 1–16 in *The Medieval Manuscript Book: Cultural Approaches*. Cambridge: Cambridge University Press, 2015.

Kamp, S. "Handschriften Lesen Lernen Im Digitalen Zeitalter." pp. 111–122 in *Kodikologie und Paläographie im digitalen Zeitalter*, edited by M. Rehbein, P. Sahle, and T. Schaßan. Norderstedt: BoD, 2009.

Kichuk, D. "Metamorphosis: Remediation in Early English Books Online (EEBO)." pp. 291–303 in *Literary and Linguistic Computing* 22, no. 3 (2007).

Kiss, F.G., and et al. "Old Light on New Media: Medieval Practices in the Digital Age." pp. 16–34 in *Digital Philology: A Journal of Medieval Cultures* 2, no. 1 (2013).

Kittler, F.A. *Gramophone, Film, Typewriter*. Translated by G. Winthrop-Young and M. Wutz. Stanford: Stanford University Press, 1999 [Or. 1986].

Kleinlogel, A. "Variants and Invariants: The Logics of Manuscript Tradition." pp. 259–268 in *Theoretical Approaches to the Transmission and Edition of Oriental Manuscripts*, edited by J. Pfeiffer and M. Kropp. Beirut: Ergon Verlag, 2007.

Kratchkovsky, I.Y. *Among Arabic Manuscripts*. Translated by T. Minorsky. Leiden: Brill, 1953 [Reprinted 2016].

Krätli, G., "Between Quandary and Squander: A Brief and Biased Inquiry into the Preservation of West African Arabic Manuscripts: The State of the Discipline," pp. 399–431 in *Book History* 19 (2016).

Kropf, E., Rodgers, J., "Collaboration in Cataloguing: Islamic Manuscripts at Michigan," pp. 17–29 in *MELA Notes* 82 (2009).

Kropf, E.C., “Will that Surrogate Do?: Reflections on Material Manuscript Literacy in the Digital Environment from Islamic Manuscripts at the University of Michigan Library,” pp. 52–70 in *Manuscript Studies* vol. 1, no. 1 (2017).

Kurz, S. *Digital Humanities: Grundlagen und Technologien für die Praxis*. Wiesbaden: Springer Vieweg, 2015.

Lanham, R.A. “The Electronic Word: Literary Study and the Digital Revolution.” pp. 265–290 in *New Literary History* 20, no. 2 (1989).

Larsson, G. *Muslims and the New Media: Historical and contemporary debates*. Farnham: Ashgate, 2011.

Leemhuis, F. “From palm leaves to the Internet.” pp. 145–62 in *The Cambridge Companion to the Qur’ān*, edited by J.D. McAuliffe. Cambridge: Cambridge University Press, 2007.

Lerer, S. “Bibliographical Theory and the Textuality of the Codex: Toward a History of the Premodern Book.” pp. 17–33 in *The Medieval Manuscript Book: Cultural Approaches*, edited by M. Van Dussen and M. Johnston, Cambridge: Cambridge University Press, 2015.

Lévi-Strauss, C. *Tristes Tropiques*. Translated by J. Russell. New York: Criterion Books, 1961.

Lindgren Leavenworth, M. “Paratextual Navigation as a Research Method: Fan Fiction Archives and Reader Instructions.” pp. 51–71 in *Research Methods for Reading Digital Data in the Digital Humanities*, edited by G. Griffin and M. Hayler. Edinburgh: Edinburgh University Press, 2016.

Lit, L.W.C. van. “Commentary and Commentary Tradition: The Basic Terms for Understanding Islamic Intellectual History.” pp. 3–26 in *MIDEO* 32 (2017).

Lit, L.W.C. van. *The World of Image in Islamic Philosophy: Ibn Sīnā, Suhrawardī, Shahrazūrī, and Beyond*. Edinburgh: Edinburgh University Press, 2017.

Lit, L.W.C. van. “Islam Felsefesi ve Bilginin Dolayımi: El Yazmaları Üzerinden Yüz Yüze Sohbet.” pp. 78–81 in *Sabah Ülkesi* 52 (2017).

Lit, L.W.C. van. “Mysterious Symbols in Islamic Philosophy.” pp. 34–39 in *Islamic World of Art* 3 (2017).

Love, H. “Early Modern Print Culture: Assessing the Models.” pp. 45–64 in *Parergon* 20, no. 1 (2003).

Lynch, C. “Authenticity and Integrity in the Digital Environment: An Exploratory Analysis of the Central Role of Trust.” pp. 32–50 in *Authenticity in a Digital Environment*. Washington: Council on Library and Information Resources, 2000.

Mahdi, M. “From the Manuscript Age to the Age of Printed Books.” pp. 127–142 in *The History of the Book in the Middle East*, edited by G. Roper. Surrey: Ashgate, 2013.

Mandell, L. *Breaking the Book: Print Humanities in the Digital Age*. Chichester: Wiley-Blackwell, 2015.

Mangen, A. “Hypertext Fiction Reading: Haptics and Immersion.” pp. 404–19 in *Journal of Research in Reading* 31, no. 4 (2008).

Mak, B. *How the Page Matters*. Toronto: University of Toronto Press, 2011.

McGann, J. “Information Technology and the Troubled Humanities.” pp. 105–21 in *TEXT Technology* 14, no. 2 (2005).

McGann, J. *A New Republic of Letters: Memory and Scholarship in the Age of Digital Reproduction*. Cambridge Mass.: Harvard University Press, 2014.

McGillivray, M. “Statistical Analysis of Digital Paleographic Data: What Can It Tell Us?” pp. 47–60 in *TEXT Technology* 1 (2005).

McGrady, D. “Textual Bodies, the Digital Surrogate, and Desire: Guillaume de Machaut’s Judgment Cycle and His Protean Corpus.” pp. 8–27 in *Digital Philology: A Journal of Medieval Cultures* 5, no. 1 (2016).

McLaughlin, T. *Reading and the Body*. New York: Palgrave Macmillan, 2015.

McLuhan, M. *Understanding Media: The Extensions of Man*. Cambridge Mass.: MIT Press, 1994 [Or. 1964].

Meinlschmidt, P., C. Kämmerer, and V. Märgner. “Thermographie—Ein Neuartiges Verfahren Zur Exakten Abnahme, Identifizierung Und Digitalen Archivierung von Wasserzeichen in Mittelalterlichen Und Frühneuzeitlichen Papierhandschriften, -Zeichnungen Und -Drucken.” pp. 209–226 in *Kodikologie Und Paläographie Im Digitalen Zeitalter 2*. Norderstedt: Books on Demand, 2010.

Merkoski, J. *Burning the Page: The Ebook Revolution and the Future of Reading*. Naperville: Sourcebooks, 2013.

Michot, Y. *Muslims under Non-Muslim Rule*. Oxford: Interface Publications, 2006.

Michot, Y. “Ibn Taymiyya’s ‘New Mardin Fatwa’. Is Genetically Modified Islam (GMI) Carcinogenic?” pp. 130–181 in *The Muslim World* 101, no. 2 (2011).

Miller, M.T., M.G. Romanov, and S.B. Savant. “Digitizing the Textual Heritage of the Premodern Islamicate World: Principles and Plans.” pp. 103–109 in *International Journal of Middle East Studies* 50, no. 1 (2018).

Monteil, V. “La Cryptographie Chez Les Maures.” pp. 1257–1264 in *Bulletin de l’Institut Français d’Afrique Noire* 13, no. 4 (1951).

Morris, E. “The Certainty of Donald Rumsfeld.” *The New York Times*. March 25, 2014.

Muir, B.J. “Innovations in Analyzing Manuscript Images and Using Them in Digital Scholarly Publications.” pp. 135–144 in *Kodikologie und Paläographie im digitalen Zeitalter*, edited by M. Rehbein, P. Sahle, and T. Schaßan. Norderstedt: BoD, 2009.

Mukhtar Umar, A., and A. Salim Mukarram (eds.). *Muʿjam al-qirāʾāt al-qurāniyya*. 8 vols. Kuwait: Dhāt al-salāsil, 1988.

Müller, J.D. “The Body of the Book: The Media Transition from Manuscript to Print.” pp. 32–44 in *Materialities of Communication*, edited by H.U. Gumbrecht and K.L. Pfeiffer, Stanford: Stanford University Press, 1994.

Muri, A. “The Grub Street Project: Imagining Futures in Scholarly Editing.” pp. 15–26 in *Online Humanities Scholarship: The Shape of Things to Come*, edited by J. McGann. Houston: Connexions, 2010.

Murray, J.D., and W. VanRyper. *Encyclopedia of Graphic File Formats*. 2nd ed. Bonn: O’Reilly & Associates, 1996.

Nemeth, T. *Arabic Type-Making in the Machine Age: The Influence of Technology on the Form of Arabic Type, 1908–1993*. Leiden: Brill, 2017.

Neuroth, H., A. Rapp, and S. Söring, eds. *TextGrid: Von Der Community—Für Die Community*. Glückstadt: Verlag Werner Hülsbusch, 2015.

Nichols, S.G. “Introduction: Philology in a Manuscript Culture.” pp. 1–10 in *Speculum* 65, no. 1 (1990).

Nichols, S.G., and N.R. Altschul. “Digital Philology: A Journal of Medieval Cultures.” pp. 1–2 in *Digital Philology: A Journal of Medieval Cultures* 1, no. 1 (2012).

Nichols, S.G. “What Is a Manuscript Culture? Technologies of the Manuscript Matrix.” pp. 34–59 in *The Medieval Manuscript Book: Cultural Approaches*, edited by M. Van Dussen and M. Johnston, Cambridge: Cambridge University Press, 2015.

Nichols, S.G., “Materialities of the Manuscript: Codex and Court Culture in Fourteenth-Century Paris,” pp. 26–58 in *Digital Philology: A Journal of Medieval Cultures*, vol. 4, no. 1 (2015).

Nolan, M. “Medieval Habit, Modern Sensation: Reading Manuscripts in the Digital Age.” pp. 465–476 in *The Chaucer Review* 47, no. 4 (2013).

Nunberg, G. “The Places of Books in the Age of Electronic Reproduction.” pp. 13–37 in *Representations* 42 (1993).

Nyhan, J. “Joint and Multi-Authored Publication Patterns in the Digital Humanities.” pp. 387–399 in *Literary and Linguistic Computing* 29, no. 3 (2014).

Oman, G. “Maṭbaʿa”, vol. VI, p. 795a in Bearman, P., Th. Bianquis, C.E. Bosworth, E. van Donzel, and W.P. Heinrichs, eds. *The Encyclopaedia of Islam*. 2nd ed. 13 vols. Leiden: Brill, 1955–2005.

Ong, W.J. *Orality and Literacy*. London: Routledge, 2002 [Or. 1982].

Ornato, E. “La Numérisation Du Patrimoine Livresque Médiéval : Avancée Décisive Ou Miroir Aux Alouettes ?” pp. 85–115 in *Kodikologie Und Paläographie Im Digitalen Zeitalter 2*, edited by F. Fischer, Chr. Fritze, and G. Vogeler. Norderstedt: BoD, 2010.

Patten, E., and J. McElligot, eds. *The Perils of Print Culture: Book, Print and Publishing History in Theory and Practice*. London: Palgrave Macmillan, 2014.

Pedersen, J. *The Arabic Book*. Translated by G. French. Princeton: Princeton University Press, 1984.

Peursen, W. van. “Text Comparison and Digital Creativity: An Introduction.” pp. 1–30 in *Text Comparison and Digital Creativity*, edited by W. van Peursen, E.D. Thoutenhoofd, and A. van der Weel. Leiden: Brill, 2010.

Pierazzo, E. *Digital Scholarly Editing: Theories, Models and Methods*. Farnham: Ashgate, 2015.

Pierazzo, E. "Textual Scholarship and Text Encoding." pp. 307–321 in *A New Companion to Digital Humanities*, edited by S. Schreibman, R. Siemens, and J. Unsworth. Oxford: Wiley-Blackwell, 2016.

Plotinus. *Plotini Opera. Enneades IV–V* [*Plotiniana Arabica Ad Codicum Fidem Anglice Vertit*]. Edited by P. Henry and H.R. Schwyzer. Translated by G. Lewis. Paris: Desclée de Brouwer et Cie, 1959.

Poirion, D. "Ecriture et Ré-Écriture Au Moyen Âge." pp. 109–118 in *Littérature* 41 (1941).

Prescott, A. "Consumers, Creators or Commentators? Problems of Audience and Mission in the Digital Humanities." pp. 61–75 in *Arts & Humanities in Higher Education* 11, no. 1–2 (2011).

Puin, G.-R. "Vowel Letters and Ortho-Epic Writing in the Qurʾān." pp. 147–90 in *New Perspectives on the Qurʾān: The Qurʾān in Its Historical Context 2*, edited by G.S. Reynolds. London: Routledge, 2011.

Rachman, T. "Writers gonna write" pp. 8–9 in *The Times Literary Supplement*, no. 5990, January 19 2018.

Rafiyenko, D. "Tracing: A Graphical-Digital Method for Restoring Damaged Manuscripts." pp. 121–135 in *Kodikologie und Paläographie im digitalen Zeitalter 4*, edited by H. Busch, F. Fischer, and P. Sahle. Norderstedt: BoD, 2017.

Rehbein, M., and Chr. Fritze. "Hands-On Teaching Digital Humanities." pp. 47–78 in *Digital Humanities Pedagogy: Practices, Principles and Politics*. Cambridge: Open Book Publishers, 2012.

Ridwan, A. *Taʾrīkh maṭbaʿa būlāq*. Cairo: Bulaq, 1953.

Riedel, D. "How Digitization Has Changed the Cataloging of Islamic Books." *Research Blog Islamic Books*, August 14, 2012.

Rieger, O.Y., "Enduring Access to Special Collections: Challenges and Opportunities for Large-Scale Digitization Initiatives," pp. 11–22 in *RBM* vol. 11, iss. 1 (2010).

Rimmer, J., C. Warwick, A. Blandford, J. Gow, and G. Buchanan. "An Examination of the Physical and the Digital Qualities of Humanities Research." pp. 1374–1392 in *Information Processing and Management* 44 (2008).

Rippin, A. "The Qur'ān on the Internet: Implications and Future Possibilities." pp. 113–26 in *Muslims and the New Information and Communication Technologies*, edited by T. Hoffmann and G. Larsson. New York: Springer, 2013.

Ritter, H., "Autographs in Turkish Libraries," pp. 63–90 in *Oriens* vol. 6, no. 1 (1953).

Robinson, F. "Technology and Religious Change: Islam and the Impact of Print." pp. 229–51 in *Modern Asian Studies* 27, no. 1 (1993).

Robinson, P. "Current Issues in Making Digital Editions of Medieval Texts—or, Do Electronic Scholarly Editions Have a Future?" *Digital Medievalist* 1 (2005).

Robinson, P. "The Digital Revolution in Scholarly Editing." pp. 181–207 in *Ars Edendi Lecture Series, Vol. IV*, edited by B. Crostini, G. Iversen, and B.M. Jensen. Stockholm: Stockholm University Press, 2016.

Roerich, N. *The Invincible*. New York: Nicholas Roerich Museum, 1994.

Romanov, M.G. "Observations of a Medieval Quantitative Historian?" pp. 462–496 in *Der Islam* 94, no. 2 (2017).

Rosenthal, F. *The Technique and Approach of Muslim Scholarship*. Rome: Pontificium Institutum Biblicum, 1947.

Ruff, C. "Scholars Criticize Academia.edu Proposal to Charge Authors for Recommendations." *The Chronicle of Higher Education*, January 29, 2016.

Ryan, D., "Aluka: digitization from Maputo to Timbuktu," pp. 29–38 in *OCLC Systems & Services: International digital library perspectives* vol. 26, iss. 1 (2010).

Sahle, P. *Digitale Editionsformen: Zum Umgang mit der Überlieferung unter der Bedingungen des Medienwandels*. 3 vols. Norderstedt: BoD, 2013.

Said, E. *Orientalism*. New York: Pantheon Books, 1978.

Schreibman, S., R. Siemens, and J. Unsworth, eds. *A New Companion to Digital Humanities.* Oxford: Wiley-Blackwell, 2016.

Sentilles, R.M. "Toiling in the Archives of Cyberspace." pp. 136–156 in *Archive Stories: Facts, Fictions, and the Writing of History*, edited by A. Burton. Durham: Duke University Press, 2005.

Shiel, P., M. Rehbein, and J. Keating. "The Ghost in the Manuscript: Hyperspectral Text Recovery and Segmentation." pp. 159–174 in *Kodikologie Und Paläographie Im Digitalen Zeitalter*. Norderstedt: Books on Demand, 2009.

Skovgaarden-Petersen, J. *Defining Islam for the Egyptian State: Muftis and fatwas of the Dar Al-ifta*. Leiden: Brill, 1997.

Smith, N. "Digital Infrastructure and the Homer Multitext Project." pp. 121–38 in *Digital Research in the Study of Classical Antiquity*, edited by G. Bodard and S. Mahony. Farnham: Ashgate, 2010.

Snow, C.P. *The Two Cultures and the Scientific Revolution*. Cambridge: Cambridge University Press, 1959.

Solth, A., R. Herzog, and M. Neumann. "A Modular Workbench for Manuscript Analysis." pp. 132–137 in *Manuscript Cultures* 7 (2013).

Stansbury, M. "The Computer and the Classification of Script." pp. 237–249 in *Kodikologie und Paläographie im digitalen Zeitalter*, edited by M. Rehbein, P. Sahle, and T. Schaßan. Norderstedt: BoD, 2009.

Stapel, R. "The Development of a Medieval Scribe." pp. 67–86 in *Kodikologie und Paläographie im digitalen Zeitalter 3*, edited by B. Assmann, J. Puhl, and P. Sahle. Norderstedt: BoD, 2015.

Stella, F. "Digital Philology, Medieval Texts, and the Corpus of Latin Rhythms, a Digital Edition of Music and Poems." pp. 223–249 in *Digital Philology and Medieval Texts*, edited by A. Ciula and F. Stella. Pisa: Pacini, 2006.

Stinson, T. "Counting Sheep: Potential Applications of DNA Analysis to the Study of Medieval Parchment Production." pp. 191–207 in *Kodikologie Und Paläographie Im Digitalen Zeitalter 2*. Norderstedt: Books on Demand, 2010.

Stokes, P. "Palaeography and Image-Processing: Some Solutions and Problems." *Digital Medievalist* 3 (2007).

Stokes, P. "Computer-Aided Palaeography, Present and Future." pp. 309–337 in *Kodikologie und Paläographie im digitalen Zeitalter*, edited by M. Rehbein, P. Sahle, and T. Schaßan. Norderstedt: BoD, 2009.

Stokes, P. "Teaching Manuscripts in the Digital Age." pp. 229–245 in *Kodikologie und Paläographie im digitalen Zeitalter 2*, edited by F. Fischer, Chr. Fritze, and G. Vogeler. Norderstedt: BoD, 2010.

Stokoe, Chr., G. Ferrario, and M. Outhwaite. "In the Shadow of Goitein: Text Mining the Cairo Genizah." pp. 29–34 in *Manuscript Cultures* 7 (2013).

Stutzmann, D. "Paléographie Statistique Pour Décrire, Identifier, Dater … Normaliser Pour Coopérer et Aller plus Loin ?" pp. 247–277 in *Kodikologie Und Paläographie Im Digitalen Zeitalter 2*, edited by F. Fischer, Chr. Fritze, and G. Vogeler. Norderstedt: BoD, 2010.

Suhrawardī, *The Philosophy of Illumination* [= Ḥikmat al-ishrāq], Translated by J. Walbridge and H. Ziai, Provo: Brigham Young University Press, 1999.

Suhrawardī, *al-Mashāriʿ*, in *Opera Metaphysica et Mystica* [= Oeuvres Philosophiques et Mystiques / Majmūʿa fī l-ḥikma al-ilāhiyya], Edited by H. Corbin, 4 vols., Orig. publ. 1945–1970., Tehran: Institut franco-iranien, 2009.

Sutherland, K. "Being Critical: Paper-Based Editing and the Digital Environment." pp. 13–26 in *Text Editing, Print and the Digital World*, edited by M. Deegan and K. Sutherland. Surrey: Ashgate, 2009.

Swanick, S., "Of making books there is no end: Islamic manuscripts on the Web," pp. 416–419 in *College and Research Libraries News* (July/August 2011).

Szpiech, R. "Cracking the Code: Reflections on Manuscripts in the Age of Digital Books." pp. 75–100 in *Digital Philology: A Journal of Medieval Cultures* 3, no. 1 (2014).

Teasdale, M.D., S. Fiddyment, J. Vnoucek, V. Mattiangeli, C. Speller, A. Binois, M. Carver, et al. "The York Gospels: A 1000-Year Biological Palimpsest." pp. 1–11 in *Royal Society Open Science* 4 (2017).

"Technical Guidelines for Digitizing Cultural Heritage Materials." Federal Agencies Digital Guidelines Initiative, 2016.

Terras, M.M. "Artefacts and Errors: Acknowledging Issues of Representation in the Digital Imagining of Ancient Texts." pp. 43–61 in *Kodikologie Und Paläographie Im Digitalen Zeitalter 2*, edited by F. Fischer, Chr. Fritze, and G. Vogeler. Norderstedt: BoD, 2010.

Terras, M. "Present, Not Voting: Digital Humanities in the Panopticon." pp. 172–90 in *Understanding Digital Humanities*, edited by D.M. Berry. New York: Palgrave Macmillan, 2012.

"The Digital Middle Ages." *Speculum* 92, no. S1 (2017).

Treharne, E. "Fleshing out the Text: The Transcendent Manuscript in the Digital Age." pp. 465–478 in *Postmedieval: A Journal of Medieval Cultural Studies* 4, no. 4 (2013).

Vismann, C. *Files*. Translated by G. Winthrop-Young. Stanford: Stanford University Press, 2008.

Volosinov, V.N. *Marxism and the Philosophy of Language*. Translated by L. Matejka and I.R. Titunik. New York: Seminar Press, 1973.

Warwick, C., M.M. Terras, P. Huntington, and N. Pappa. "If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data." pp. 85–102 in *Literary and Linguistic Computing* 23, no. 1 (2008).

Weel, A. van der. "New Mediums: New Perspectives on Knowledge Production." pp. 253–268 in *Text Comparison and Digital Creativity*, edited by W. van Peursen, E.D. Thoutenhoofd, and A. van der Weel. Leiden: Brill, 2010.

Weiss, A. *Using Massive Digital Libraries*. Chicago: ALA TechSource, 2014.

Witkam, J.J., "Establishing the Stemma: Fact or Fiction?", pp. 88–101 in *Manuscripts of the Middle East* 3 (1988).

Witkam, J.J. *Inventory of the Oriental Manuscripts in Leiden University Library*. 28 vols. Leiden: Ter Lugt Press, 2007.

Wittgenstein, L. *Tractatus Logico-Philosophicus*. Translated by D.F. Pears and B.F. McGuinness. London: Routledge, 2001 [1921].

Wogan-Browne, J., N. Watson, A. Taylor, and R. Evans, eds. *The Idea of the Vernacular: An Anthology of Middle English Literary Theory 1280–1520*. Exeter: University of Exeter Press, 1999.

Wolf, L., N. Dershowitz, L. Potikha, T. German, R. Shweka, and Y. Choueka. "Automatic Palaeographic Exploration of Genizah Manuscripts." pp. 157–179 in *Kodikologie und Paläographie im digitalen Zeitalter 2*, edited by F. Fischer, Chr. Fritze, and G. Vogeler. Norderstedt: BoD, 2010.

Zaynab, A.N., A. Abrizah, and M.R. Hilmi. "What a Digital Library of Malay Manuscripts Should Support: An Exploratory Needs Analysis." pp. 275–289 in *Libri* 59 (2009).

Zorich, D.M. "Digital Humanities Centers: Loci for Digital Scholarship." pp. 70–78 in *Working Together or Apart: Promoting the Next Generation of Digital Scholarship*. Washington, D.C.: Council on Library and Information Resources, 2009.

Zulkifli, Z. "A Collaborative E-Workspace for Digital Library of Malay Manuscripts." pp. 368–372 in *International Journal of Information and Education Technology* 4, no. 4 (2014).

Zundert, J.J. van. "By Way of Conclusion: Truly Scholarly, Digital, and Innovative Editions?" pp. 335–346 in *Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches*, edited by T. Andrews and C. Macé. Turnhout: Brepols, 2014.

Index of Persons

Index of Subjects