A close-up of a book

Description automatically generated

A nice example of manuscript complexity and the often emergent nature of the whole, which may or may not be readily expressed in datafied terms: an early 19th century (probably) copy of [*Tārīkh-i ilāhī haz̤rat Akbar padishah* (Lewis O 45)](https://openn.library.upenn.edu/Data/0023/html/lewis_o_045.html)

{: .figcaption }

The following is a distillation of remarks I had prepared for a digital humanities roundtable at last week’s annual Middle Eat Studies Association meeting in Montreal; due to suddenly getting quite sick at the last minute I was forced to cancel my travel plans and so was unable to participate. In lieu of that participation I’ve reworked that material into essay form, with the goal of thinking out loud regarding the intersection of manuscript studies, machine learning, ‘datafication,’ and digital humanities as a field and discipline, among other things.

\_\_\_\_\_\_\_\_\_\_\_

i.

Let me preface my remarks by noting that I feel somewhat underqualified in writing this essay: all of the discrete components in this discussion—manuscript studies, the history of textual technologies, book history, computational methods, machine learning, and so forth—are things I’ve come to relatively recently, not having particularly featured in my graduate studies. That said, I’d like to think that while I cannot claim to be master of any of the relevant fields of study involved here, I have spent enough time around these topics so as to have something that is, I hope, interesting and useful.

There are a lot—perhaps too many—things that get included under the umbrella term ‘digital humanities,’ with some definitions of ‘digital humanist’ encompassing anyone who uses a search engine and a word processor, to others which restrict it to individuals capable of coding and doing computational projects more or less independently. I prefer a definition that leans towards the more catholic end, with the stipulation that one is engaged in digital humanities when one is not only using digital tools and methods, but does so in a critical and self-aware fashion, informed by knowledge of the shape and structure of digital systems, their affordances and limitations, and cognizance of the implications of the interplay of the digital and non-digital. The ability to code is in this understanding not necessary, though some knowledge of how coding works and of the ‘invisible’ infrastructure of the digital world is a prerequisite, preferably with critical reflection on the epistemic, phenomenological, and other implications of that digital infrastructure and secondary effects.

One of the commonalities to all definitions and practices of digital humanities is the generation and use of bodies of *data* at scale: indeed, one of the most powerful, and potentially fraught, affordances of digital technology is the possibility of commanding very large corpora of electronic texts (and other sorts of datasets in other contexts), operating at scales that would be impossible for non-digitally assisted humans. While computational methods do not require *per se* large electronic corpora—one can perform computational and other digitally-assisted operations upon a text of a few folios worth, after all—perhaps the greatest excitement and promise in our field comes from being able to compile and render legible corpora at prodigious scales. Even prior to questions to be asked of large corpora is the task and promise of bringing the disparate elements together and creating or reproducing sufficient metadata to make such large corpora navigable in some way.

Things become especially interesting when we begin talking about the ‘datafication’ of manuscript texts once the digitized exemplars have been assembled into corpora. First off, I want to make a couple of assumptions that while I think realistic are not as yet set in stone, as we are still refining the technology in question: I want us to imagine that at some point in, probably, the near future, we have sufficiently perfected handwritten text recognition, segmentation, and region analysis such that we can in fact automatically generate very large electronic textual corpora using manuscripts as our basis. I’m going to take a further step and suppose that we will obtain a decently high level of transcription accuracy, enough that we can apply computational methods, employ good quality internal search, and use the raw results as the starting point for refined transcriptions and editions; these are all likely results, even if it is unlikely that we will ever obtain identical results to optical character recognition for typographic print, or that the generation of ‘proper’ editions of manuscript texts will ever be fully automated. Still, let us assume that the effective datafication of manuscripts will be possible, sooner probably rather than later, and that in the near future we will have a vast corpus of electronic texts derived from manuscripts.

From such a starting point, we can then ask: how exactly should we think about electronic texts, about bodies of data, derived from digitized manuscripts? What special concerns, limitations, and possibilities might such a corpus hold vis-à-vis other forms of textual corpora? How do the affordances and informational deposits internal to manuscripts *qua* manuscripts translate, or not translate, into ‘datafied’ electronic formats? How might quantifiable, computational methods work within such a corpus? What is gained in such a scenario, and what is potentially lost? How can gains be amplified, and losses mitigated—or at least registered?

ii.

It might be helpful to begin answering such questions by thinking about the points of continuity and connection between the internal development and shape of the textual traditions we study, on the one hand, and contemporary ways of thinking about, generating, and using texts and information more widely. In the future I hope to explore these ‘correspondences,’ not so much genealogical as cultural evolutionarily convergent, between manuscript culture and digital culture, in particular the ‘computational’ undergirding that is visible in some aspects of manuscript culture and is of course so crucial in our own world. For the moment however, let us think a bit about issues of data, scale, legibility, and paratextual and metadata apparatuses, places in which we can see both overlap and divergence and potential friction.

‘Datafication’ is not in a certain sense at least particularly new when it comes to Islamicate book cultures, given the veritable ‘economies of scale’ we see in some aspects of textual production and reproduction in the premodern world. In fact one of the initial animating concerns behind OpenITI has been the desire to make sense of book production in the premodern Islamicate, grappling with a question that surely anyone who has spent even a little time around premodern Islamicate textual traditions has broached: how *did* these authors create such massive multi-volume works? Think of *tafsīr* and *tārīkh* works whose modern printed editions suggest the need for a wheelbarrow to cart them around; one grows weary just thinking of the labor of compiling them. It has become clear that these works were built upon a certain utilitarian approach to vast bodies of text, removing and remixing extensive sections from older works, rendering those older works something like a corpus for text mining, albeit unassisted by any large language models and instead only the scribal eye and pen. When viewed at scale, thanks to the development of both large corpora and the tools and visual outputs providing analysis of this data, we can begin to see large-scale, long chronology patterns and dynamics. Here there is a useful symmetry between premodern forms of textual production and the affordances of contemporary technology.

We can also see the transformation over time of the paratextual apparatuses through which premodern reader and writers navigated often quite large bodies of texts or individual texts of prodigious size; searchability, findability, and detection of connections, authorities, and so forth, were all concerns whose material traces remain in the manuscripts that have come down to us. The methods changed over time: we can think of the rise and increasing purchase of the table of contents, for instance, as a seemingly simply finding device that nonetheless could make quite a difference in how a manuscript was used and by whom (and which might tell us a great deal about the theory of reading and information/communication implicit in its production and use). In fact one potential area of future research of a quantifiable sort might well be tracing, at scale, the development, articulation, and spread of such paratextual devices and the deployment of ‘metadata’ within texts in reference to other works and corpora. Here we are helped by a de facto correspondence between premodern and contemporary approaches, even if in this case we cannot descry direct genealogical connection, and the technological and theoretical parameters governing the two milieus are otherwise quite different. The basic parameters have a congruence across time and media, despite the many dramatic changes and expansions of the digital age.

iii.

There are, I think, some special considerations when it comes to manuscript culture and the electronic rendering of the contained texts, taken as discrete and (at least partially) detachable objects, capable of reduction to an encoded electronic text absent its material context (and the context of those semantic elements which resist reduction or reproduction in a data environment). One of the most salient areas in which operations of scale and a considerable degree of datafication pose both great possible rewards and assured challenges is the question of *textual variability* in manuscripts, a question that at the outer edges will tend to challenge our typographic print-centered assumptions about what makes a discrete text after all.

When we are producing corpora of electronic texts derived from printed sources, we can more or less safely assume that we are dealing with discretely bounded texts, for which it makes sense to say that, say, *Tafsīr al-Fulānī* consists of a quite stable number of words, possibly spread across several editions but with variability minor on the whole, essentially ‘noise.’ Here datafication and resulting analysis builds off of the prior transformation of textuality in the movement from manuscript to typographic print, which generated a whole host of other changes to book culture and practices of textuality. For manuscript texts we are dealing with rather different textual creatures: if there are twenty discrete manuscript instances of this *tafsīr* odds are good that they range widely in word count, in order, in number of volumes present, and so forth, to the point that for some texts it is hard to say just what precisely really constitutes the discrete text across versions.

Assuming, then, that we are able to assemble large corpora of electronic texts derived from manuscript exemplars—the precise language here itself becomes tricky, in fact—we might then ask what exactly we are looking at when we do computational analysis, and how we might segregate, label, subdivide, and then tackle such data so as to reflect the internal complexity and frequent ambiguity internal to it, at base the question of *Just what is a ‘text’ in this context anyway?* This complexity is certainly a major challenge, but it also presents ample opportunity for thinking at scales and with varying resolutions not possible through traditional ‘analog’ means. We have begun to make initial stabs—with a stress on the initiality of our efforts—at using digital methods to explore manuscript text variability; there is much more that needs to be done, though at least here, as with text reuse analysis (which itself is likely to take new directions once manuscript derived data can be applied), the affordances of the digital and of datafication track well into the shape of the premodern material and provide ways to reveal patterns and dynamics really only fully visible at scale and distance.

iv.

Perhaps the most important difference between textual datafication drawing upon typographically printed works and datafication drawing upon manuscripts (with lithography its own intermediate case!) is that while the materiality of printed books is not unimportant, there is a closer relationship between the action of turning typographic print into electronic text and hence into data: typography was perhaps the premier pioneer technology in applying energy and technique to the simplification, amplification, and mass reproduction of a previously quite complex and materially robust entity, namely manuscript texts. Each letterform is created on the page using a vastly reproducible single entity, a piece of material type; the letters are held within a form, marking off the margins as not-text, restraining any textual apparatus in a precisely calibrated relationship to the main text. There is a hygienic neatness and a stability to the typographic page, which would in time figure into the often dramatic transformations textuality has undergone in the transition from manuscript to print.

To be sure as in anything in the sublunary realm things are not universally stable and clean, but in the great scheme of things that is rather immaterial (pun intended). Typographic print, not entirely due to endogenous material constraints (one can do all kinds of creative things with type, after all) but as much due to the larger social and economic relationships and requirements, the affordances permitted by the human structures in which the technology itself unfolded, has largely done away with the play and rampant vegetational productivity of the manuscript page. For if typographic print inclined towards an easily and universally legible textual-visual monoculture (viz., *Mono*type!), manuscript cultures have historically usually been polycultures of the page, with some Islamicate genres excelling in the degree to which texts in multiple hands and often distributed over years or centuries could run riot on the page, interpenetrating the main text or even obviating the need or the logic of identifying a main text at all.

Digital textuality is *phylogenetically* at least a descendant of typographic print, and has heightened many of its particularities: while it is not true that code, for instance, obviates human creativity or particularity, it is also the case that one cannot be eccentric or cavalier when working in code; slight errors in the command line can cause big headaches. Regimes of normalization are vital, and output in turn tends towards a visual and semantic sameness, relationships and particularities registered in number, with binary code at the base of everything. While it remains possible to indicate the spatial relationship of one block of text on a page to another block or isolated lines, the indication of such relationships is different from the visual realization conveyed on the original manuscript page. We are talking about a multi-layered shift in the primary phenomenology of reading and otherwise interacting with the text. Yet even if digital texts and data are the evolutionary descendants of print, by a sort of convergent process they have also developed certain characteristics of the manuscript text, or, more appropriately, the culture and norms within which those texts were created and reproduced. The fluidity of textual production and reproduction has been greatly accelerated by the digital; with texts becoming more and more detached from a consistent or coherent human authorial voice, fragments and pieces of texts moving in and out of various modalities of us, from the raw data of LLM to the strange facsimile nature of the memetic copypasta.

The datafication of manuscripts, whether through the production and analysis of metadata or through the registering and extraction of text, runs the danger of neglecting those elements of manuscripts that are either non-semantic or which are not legible to existing machine learning based models and approaches. The spatial arrangement of text on the page, the variation in hands, scripts, and the like, the marks of use and wear and tear—these are all components of a manuscript that potentially convey information and shape our interpretations of the text under consideration. It is conceivable that most or all of these components could be themselves ‘datafied’ and analyzed at scale, even if at present for our corpora at least we lack such tools and approaches.

What is crucial at this juncture is understanding all of these elements and their particularities, appreciating that as in so much else in our world there is an emergent object resultant from all of these discrete elements that is not reducible just to the raw numerical combination of those elements. All that which is disaggregated for purposes of legibility and analysis must ultimately be brought back together: in other words, even critically and carefully undertaken use of data must always coexist with other forms of interaction and analysis, with no single approach or focus on one particular facet an end-all in itself.