METADATA IN THE MARGINS: RESHAPING ARCHIVES AS DATA THROUGH EARLY MODERN MARGINALIA

by

CHANTAL BROUSSEAU, B.A. Honours

A research essay submitted to Carleton University in fulfillment of the requirements for the course HIST 5908, as credit toward the degree of Master of Arts in History – with Specialization in Data Science

Department of History
Carleton University
Ottawa, Canada

September 17, 2023

Abstract

In this research essay, along with its associated application and case study, I delve into the transformative impact of digitized cultural heritage collections on emerging research methodologies. I have identified key issues which digitized cultural heritage collections and by extension, institutions, face during this time of rapid technological development, specifically in the realm of machine learning. I address this through the development of an application to situate the data which machine learning models are built on, and further demonstrate the positive impact machine learning has the potential to make on the study of history through a case study on the detection of marginalia in the National Library of Scotland's digitized collection of largely early modern chapbooks. My research ultimately showcases how machine learning, when applied mindfully and with a dedicated focus on ethical considerations, can bring about beneficial changes in the field of history research.

Acknowledgements

I would first and foremostly like to express my sincere gratitude to my supervisor, Shawn Graham, for his invaluable guidance, unwavering support, and insightful feedback throughout the course of this research. His expertise and mentorship have been instrumental in shaping the direction and final product of this thesis, and I am truly grateful for the time and effort he dedicated to helping me succeed in my entire time at Carleton.

I am also deeply appreciative of my family and friends, who provided me with the space, encouragement, and understanding I needed during this challenging process. Their consistent belief in me, coupled with the occasional insistence on taking breaks and garden-fresh vegetables, has been a source of strength and motivation, and I am fortunate to have such a remarkable support system.

Finally, a special note of thanks goes to my cats, Peaches and Basil. Peaches, by sitting on me for extended periods of time, often served as a peculiar yet effective reminder to remain focused and committed to my work. Feral kitten Basil, who appeared in the midst of my MRE process on our back porch, was a forceful reminder of the balance between life and studies that I often needed to be reminded of. Their companionship during late-night writing sessions brought a sense of comfort and added a touch of warmth to the occasional solitude of research.

Table of Contents

List of Figures

List of Tables

List of Appendices

Note on this Work

This essay was originally composed in the form of a website. A link to this website, along with a link to all other technical components of this work can be found in **Appendix A.** On the website, the technical information that makes up **Appendix C** is integrated into the main text to offer digital readers an article with more of a technical slant. However, there are also interactive graphics on the website, so I suggest all readers pay the website a visit, even if just to look rather than read.

Introduction

While studying texts of the past, it is not unusual to stumble upon evidence that a document had a 'life' before coming to reside in the archive where it now is found: numerous notes throughout the text at the bottom of the page made by the original purchaser, corrections scrawled in rough handwriting from a child using the document as reading practice, or perhaps there is even an initial on the title page left from the document's first foray into the archive. Indeed, these markings known formally as marginalia not only served a purpose to those who created them, but they also serve a purpose for historians through how they may situate a text within its history and allow for a glimpse into the public and private lives of the annotator.

Despite the insights which marginalia can offer historians of reading, the book, and beyond, the study of marginalia proves to be challenging due to its inherent nature as an element residing in the margins, often scorned or overlooked during the archival process. Most studies of marginalia focus on tracing select annotators or on small collections, at least partially due to the difficulty finding marginalia across larger collections when these annotations are neither abundant nor conspicuous. It is this issue of discoverability which my research project addresses in the form of a case study, demonstrating the usage of an application I built to identify marginalia contextualized within the ongoing conversations surrounding the reconfiguration of digitised cultural heritage collections as data.

Over the last thirty years, historical study has been revolutionized by the rapid emergence of digitised resources which have become widely available to the public, yet techniques which take advantage of the unique digital affordances of such representations are still being developed. A key component of these digitised materials is the metadata which situates them. This metadata describes the object both as a unique digital entity and as the original object it represents. Metadata within digital archives have been used by scholars such as Ryan Cordell to demonstrate the political and social contexts that inform such corpora of materials. Yet there are other forms of metadata generated when scholars use these digitised resources, particularly when they are adapted for use in a data-forward project, that remain unaccounted for.

My MRE project develops an approach to capture this missing metadata. I build, and critically situate, an image annotation application for identifying notable material features from digitised documents, with the focus being placed on marginalia composed by readers of these documents. The tool functions both manually and automatically, at scale for one document or multiple. Drawing on my experience publishing in The *Programming Historian*, my MRE designs, tests, and describes the tool in such a way that other scholars can immediately deploy it for their own research.[1] Tools used in research are theory-laden in that there are always choices to be made; my MRE situates these choices in such a way that the scholar who uses the tool will understand the

---

[1] Chantal Brousseau, "Interrogating a National Narrative with GPT-2," *Programming Historian*, October 2022, https://doi.org/https://doi.org/10.46430/phen0104.

consequences for their own research, and make this step of the process more transparent to those consuming the output of their research.

Machine learning is a branch of artificial intelligence concerned with creating mathematical algorithms or 'formulas' that can be said to 'learn' through exposure to data, and thus improve automatically the more it is used.[2] This process results in the creation of a model, or abstraction of the patterns in the information the machine has seen, which is then able to make predictions or decisions about new data it has not previously seen. Discovery of marginalia across expansive collections is an exemplary case of when an object detection model should be used; as the name implies, this is a type of machine learning model designed to identify objects in an image. Yet the usage of machine learning has ethical implications both broadly and in particular when used with cultural heritage collections. Questions surrounding who gets to decide which information is used for training, whose culture becomes the standard, whose voices are left out, and who profits from the work all come to the forefront when using collections as data. Machine learning is a computationally intensive task, meaning that to use it, there are always costs that must be considered when using these methods. Environmental costs to power the technology required for machine learning, and financial costs to access this technology which in turn erect barriers to entry and limit who can contribute to this

[2] Brousseau, "Interrogating a National Narrative with GPT-2".

research area, are two significant concerns within discussions of machine learning usage across all disciplines at present with the rush towards creating even more vast models.[3]

The process of teaching machine learning algorithms, formally known as *training*, requires massive amounts of data to the extent that the metadata is often overlooked or omitted due to the challenge of managing and understanding such a vast quantity of data. But to truly understand the influences and limitations of a machine learning model, it is crucial to fully know the data it is built on and understand these details apart of its initial construction. For meaningful results using machine learning, it is essential to examine the input that shaped the model through understanding those who created it, what it meant to them, where it resides in both a temporal and tangible sense, and its material context. This is of particular importance when using cultural heritage collections for the purpose of machine learning, as there is the risk of models that are trained using these collections replicating the epistemologies, injustices, and anxieties exemplified by previous institutional orders and hierarchies of power.[4]

My MRE addresses this discourse through using my image annotation application and the surrounding workflow to prepare a selection of digitised archival texts which feature handwritten marginalia from the early modern period to be used as training data for an object detection model. Image annotators in general serve as tools for generating

---

[3] Emily M. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event Canada: ACM, 2021), 619, https://doi.org/10.1145/3442188.3445922.

[4] Nanna Bonde Thylstrup et al., eds., *Uncertain Archives: Critical Keywords for Big Data*, 2021, 4, https://doi.org/10.7551/mitpress/12236.001.0001.

training datasets when using image-based machine learning techniques. Their primary functionality is enabling researchers to annotate images, creating examples for the machine to learn what features in an image are considered important. Both this image annotator and workflow can be adapted to other corpora of materials beyond marginalia. This essay contextualizes the necessity for a tool such as this and the scholarly issues around its design and construction, by applying the model trained with the tool to the collection of chapbooks provided by the National Library of Scotland (NLS).[5]

These pocket-sized pieces of reading material were printed on a single sheet then folded into booklets of 8, 12, 16 and 24 pages, continuously produced in this manner from the 17th to 19th century.[6] The subject matter of chapbooks was diverse, with sermons of covenanting ministers, prophecies, last words of murderers, and biographies of famous people of the time such as Wallace, Napoleon, and Nelson. These were interspersed with works of humour, fairy tales, and poetry, not to mention manuals of instruction and almanacs. It has been estimated that around two thirds of chapbooks contain songs and poems, often under the title garlands.

Chapbook printers frequently utilized worn and broken type purchased second-hand which naturally produced rough and unrefined prints; likewise, the woodcuts used to decorate chapbooks were also cycled and reused in print, and often were not at all related to the text they were present in. Chapbooks were sold on streets and at fairs for a

---

[5] National Library of Scotland, "Chapbooks Printed in Scotland" (National Library of Scotland), accessed May 30, 2023, https://doi.org/10.34812/VB2S-9G58.

[6] Scotland, "Chapbooks Printed in Scotland".

penny a time by pedlars dubbed ‘chapmen’, a term that is related to the word ‘cheap’ but likely also related to the Anglo-Saxon ‘ceapian’, meaning to barter, buy and sell. Individuals could also buy them directly from printing shops, although one of the features of chapbooks was the proliferation of provincial imprints with places such as Fintray, Falkirk and Inveraray being a common home to cheap print shops. Chapmen were supported by running stationers to make chapbooks, alongside broadsides, the most popular reading material for the masses during the latter half of the early modern period. Chapbooks gradually disappeared in the mid 19th century due to both the rapidly increasing amount of cheap printed content available and the rise of Victorian morality which considered many chapbook publications as crude and profane. As a widely available and affordable form of entertainment and paper, chapbooks offer great potential as sites of early modern marginalia.

In creating a model specifically designed to extract marginalia, I will not only demonstrate how an image annotator such as the one I created can be used to consolidate metadata and training data, but also demonstrate how this form of image-centered machine learning can facilitate the large-scale study of reader habits and observations across collections of readers who wrote in their books during the Early Modern period. The MRE will conclude with a reflection on notable marginalia found within the NLS chapbook collection that can be used as a base for future study, as well as the historiographical impact that using this tool might have in the context of book history and more generally, as the study of history becomes increasingly digital and joins the conversations about transparency and accessible data within the humanities.

Marginalia in the Archive

The Historiography of Marginalia

Marginalia have been recognized as significant by scholars since as early as the 19th century, who used them as a means to construct the history of earlier scholars in more detail and to expand upon their published works. Particularly of note from this time was the work of literary scholar George Charles Moore Smith. In 1913 he published a book that aimed to "illustrate the life, character, and opinions" of early modern writer Gabriel Harvey using his "unpublished materials", citing Harvey's vast array of marginalia in this process of extending existing knowledge.[7] More current works tend to focus on the materiality of marginalia, using a combination of their content alongside their placement and material features of the book itself to analyze how early modern readers interacted with and used literature.[8]

One of the earlier modern seminal works in the history of reading, ""Studied for Action": How Gabriel Harvey Read His Livy" by historians Lisa Jardine and Anthony Grafton, expands further on the notes produced by Gabriel Harvey to establish marginalia as an intellectual method through microscopic analysis of his annotations inserted into Livy's ancient history of Rome.[9] The article highlights Harvey's active and purposeful

---

[7] Gabriel Harvey and George Charles Moore Smith, *Gabriel Harvey's Marginalia* (Stratford-upon-Avon: Shakespeare Head Press, 1913).

[8] Katherine Acheson, ed., *Early Modern English Marginalia* (New York: Routledge, 2019), 3, https://doi.org/10.4324/9781315228815.

[9] Lisa Jardine and Anthony Grafton, ""Studied for Action": How Gabriel Harvey Read His Livy," *Past & Present*, no. 129 (1990): 36, https://www.jstor.org/stable/650933.

approach to reading, characterized by his meticulous marginalia and extensive annotations. Through an analysis of Harvey's notes and markings, Jardine and Grafton reveal how Harvey used Livy's texts as a foundation for his own intellectual pursuits, employing them to shape his ideas, compositions, and refine his own writings.[10] Harvey's annotations not only offer insights into his personal interpretations of Livy's works but also demonstrate his scholarly engagement and his desire to apply the lessons learned from ancient history to contemporary political and cultural contexts. By examining Harvey's reading practices and how he interacted with Livy's texts, the article illustrates the active role of early modern scholars in constructing knowledge and engaging with classical literature for practical and intellectual purposes.

Moving beyond the study of past scholars, historian Heidi Brayman Hackel established the foundations of marginalia "as records of reading motivated by cultural, social, theological, and personal inclinations" through shifting the focus from scholars such as Harvey to the more casual reader.[11] In this defining work on the history of reading, *Reading Material in Early Modern England* seeks to "delineate the asymmetries of early modern English literacies and reading habits and to expand the category of readers to include a greater variety of English people"; Brayman Hackel accomplishes this by offering a comprehensive examination of the social, cultural, and intellectual contexts in which reading took place, providing insights into the reading practices of different members of society, including women, artisans, and the elite through the

---

[10] Jardine and Grafton, ""Studied for Action"," 59.

[11] Acheson, *Early Modern English Marginalia*, 15.

analysis of annotated texts in multiple formats.[12] Through extensive research and select case studies ranging from the library of Lady Anne Clifford to unspecificed annotators present in copies of Philip Sidney's *Arcadia*, Brayman Hackel delves into the materiality of texts, investigating the physical aspects of books and how they influenced reading experiences. She explores the use of illustrations, typography, and paratextual elements, demonstrating how these features shaped readers' interactions with the texts. Through further investigation into print culture, the emergence of the printing press, and the significance of libraries, bookshops, and private collections in shaping access to and availability of reading materials, Brayman Hackel also addresses the impact of censorship and the control of reading materials by established institutions such as Oxford's Bodleian Library and the authorities.[13] By extension, Brayman Hackel also examines the purposes of reading in early modern England, ranging from religious and moral instruction to entertainment and leisure. Through this exploration, she counters the fiction of "a singular ideal reader ungrounded in place or time" and instead creates "a portrait of an early modern 'gentle reader' alongside fragmentary glimpses of multiple readers at single moments in their reading lives."[14] Through emphasis on the importance of reading as a cultural and social practice and its impact on knowledge production, relationships, and the formation of early modern English society, *Reading Material in Early Modern England* contributes to our understanding of the broader historical and cultural

---

[12] Heidi Brayman Hackel, *Reading Material in Early Modern England: Print, Gender, and Literacy* (Cambridge, U.K.; New York: Cambridge University Press, 2005), 257.

[13] Brayman Hackel, *Reading Material in Early Modern England*, 84, 99.

[14] Brayman Hackel, *Reading Material in Early Modern England*, 257.

significance of reading during this period, offering valuable insights into the ways in which texts were consumed, understood, and used in early modern English society.

Closely following Brayman Hackel's publication was the release of *Used Books: Marking Readers in Renaissance England* by William H. Sherman; Sherman's work diverged from the case study model Brayman Hackel follows common in the history of reading, in that he chose to focus on *collections* of annotated books for his case studies rather than on individual readers or books. Through the isolated traces upon which his book rests, Sherman's analyses yield "some larger patterns and a more systematic sense of how a wider group of readers used a wider range of books than in previous accounts of pre-modern marginalia."[15] By analyzing the marginalia of readers who left behind "substantial annotations", Sherman uncovers the diverse motivations behind readers' interactions with books, how these marks provide insights into the readers' responses, interpretations, and both private and public connections with the text. Like Brayman Hackel, Sherman explores how these markings can reveal readers' social identities, scholarly disciplines, and cultural values. However, unlike Brayman Hackel, Sherman focuses more of his analysis not only on content of the marginalia, but the materiality of the text which this marginalia is present in, "putting books alongside the other objects…to reconstruct the material, mental, and cultural worlds of our forebears."[16] Throughout the book, he pays close attention to "patterns of use" such as bindings,

---

[15] William H. Sherman, *Used Books: Marking Readers in Renaissance England* (Philadelphia, PA: University of Pennsylvania Press, 2008), https://www.jstor.org/stable/j.ctt3fhgzw, xii.

[16] Sherman, *Used Books*, xiv.

repairs, and other signs of wear, which offer further clues about the history of ownership, as well as early modern reading habits and the networks of exchange and circulation that facilitated the dissemination of ideas through books. Further, Sherman considered how the trade of "used" books functioned, the availability and affordability of texts, and the significance of book ownership in the cultural and intellectual life of the time, concluding that "Renaissance readers have much to teach us not only about the uses of books in the past but also about attitudes toward books where the past meets the present."[17]

Uniting the work of marginalia as records of reading by Brayman Hackel and of material function by Sherman is Stephen Orgel, whose *The Reader in the Book: A Study of Spaces and Traces* examines individual acts of reading by examining books in which the text and marginalia are "in intense communication with each other, glossing, correcting, reminding, emphasizing, arguing — cases in which reading constitutes an active and sometimes adversarial engagement with the book."[18] However, in tandem with this Orgel focuses more intensely on the book's materiality than Brayman Hackel, here echoing Sherman's work, exploring the physical and conceptual spaces created within books and the traces left there by readers, offering an analysis of how reading practices have evolved across the early modern period. Drawing both upon works that are considered early modern literary classics in the present as well as books that were considered classics in their own time, Orgel delves into the concept of reading as writing,

---

[17] Sherman, *Used Books*, 151.

[18] Stephen Orgel, *The Reader in the Book: A Study of Spaces and Traces* (Oxford, UK: Oxford University Press, Incorporated, 2015), 24, http://ebookcentral.proquest.com/lib/oculcarleton-ebooks/detail.action?docID=4310757.

the way readers absorb the texts they annotate. He explores the physicality of books, including their size, format, and design, and how these factors shape the reading and by extension annotating experience. Like Brayman Hackel, *The Reader in the Book* delves into the social and cultural aspects of reading, examining how readers' identities, backgrounds, and personal experiences influence their engagement with literature. Each chapter of Orgel's work considers factors such as gender, class, and education, highlighting the diverse ways in which readers approached and unravelled texts. He joins the discussion of materiality and content through understanding how the book itself was conceptualised during the early modern period. In the shift from manuscript to print culture, the book became not simply a text but a place and property, and by extension of this the goal of printing was not exact replication of an original text but dissemination.[19] Woven throughout Orgel's book is also a timeline of the historical evolution of reading practices, from manuscript culture to the advent of printing and the digital age. Orgel ultimately offers comprehensive analysis of how changes in reading technologies and the dissemination of texts have shaped the book and reader's role over time.

Although not explicitly dealing with the study of marginalia as previously mentioned works have, Juliet Fleming's work in her book *Cultural Graphology: Writing after Derrida* focuses on material affordances particular to the book and the pen. Her framing of the medium is considered key in understanding the annotation of texts as not just evidence of an active reader, but as an act of writing "which is material, which has the power to invent things (including selves), and which exists at the intersection of

---

[19] Orgel, *The Reader in the Book*, 5, 10.

generic norms and technological affordances."[20] Fleming positions her work by expanding on the notion of "cultural graphology" that Jacques Derrida loosely proposes in his work, *Grammatology*, which examines the relationship between writing practices, materiality, and cultural contexts.[21] Fleming engages with Derrida's deconstructionist approach and expands upon his ideas to analyze how writing functions as a cultural and social practice through the lens of the writing culture in early modern England, particularly as it came to be influenced by the commercial development of print.[22] She investigates the materiality of writing, including handwriting, typography, and inscriptions, to unveil the hidden meanings and cultural significance embedded within written texts. By examining the dynamics of writing within cultural contexts, Fleming challenges the idea that writing is a neutral tool for communication. She explores how writing practices are shaped by cultural and historical factors, emphasizing the multiplicity of meanings and interpretations that emerge from written texts.

Especially relevant to the discussion of marginalia is Fleming's conceptualization of the "renaissance collage" which describes the reading undertaken in this period with scissors and knives, through the cutting of the page and associated processes of sewing, stitching, gluing, and filing.[23] She categorises the act of cutting books for the purposes of:

---

[20] Acheson, *Early Modern English Marginalia*, 4.

[21] Jacques Derrida, *Of Grammatology*, trans. Gayatri Chakravorty Spivak, Corrected ed (Baltimore: Johns Hopkins University Press, 1997), 87.

[22] Juliet Fleming, *Cultural Graphology: Writing After Derrida* (University of Chicago Press, 2016), 28.

[23] Acheson, *Early Modern English Marginalia*, 35.

> Remov[ing] proscribed or offensive material from religious texts and learning materials, to obviate the labour of copying in producing commonplace books and other compilations, to reformat texts in order to rationalize the material they contained, to provide room for marginal or other commentaries, to add other material to and thereby expand a given text, to organize their own researches, and to illustrate or embellish presentation and other manuscripts with motifs cut from printed sources.24

Thus establishing this form of readers' traces as not just the organization of written information, but also as an act of writing. For Fleming, "the cut opens, gathers and sorts; it shapes the present and introduces the future" akin to how historians of reading posit marginalia is used, as seen in Jardine and Grafton's understanding of Harvey's political notes on a text of ancient history, or Sherman's notion of private and public connections within annotations. Annotation, like cutting, is not destructive, but rather a means to grow the work being interacted with. Through her exploration of cultural graphology, Fleming invites readers, and in particular, historians of reading and the book, to critically engage with the complexities of writing as a cultural practice. She highlights the significance of materiality, historical context, and philosophical underpinnings in understanding the role of writing in shaping and reflecting cultural and social dynamics.

One of the most recent works published on the study of marginalia which has heavily shaped my understanding of it is the book *Early Modern English Marginalia*, a series of articles compiled and edited by early modern English language and literature scholar Katherine Acheson released in 2019. Part of the *Material Readings in Early Modern Culture* series, it delves into both the content and material forms of early modern

---

[24] Fleming, *Cultural Graphology*, 99.

texts, considering marginalia as both a distinct entity for scholarly analysis and as a lens through which early modern history can be interpreted. The book is divided into three parts, each representing a way in which the presence of marginalia inserts itself into different forms of research related to the history of reading and the book. The first section titled "Materialities" explores the promise provided by the intersection between material histories of the book and the turn to writing in marginalia studies.[25] It discusses the shift from physical books to electronic ones, raising questions about annotatability and the role of early modern printed books played in this. The section includes chapters on paper production, the agency of marginalia in shaping printed works, and the consideration of object marks as marginalia. It also delves into how women writers used writing, including marginalia, to create various spaces for themselves, both physical and symbolic, within the context of early modern society. The second section, "Selves", approaches books and their margins as spaces for various forms of life-writing, where individuals inscribe their personal experiences, beliefs, and identities. It examines how early modern readers used marginalia to engage with religious, political, and personal matters, demonstrating how doing so shaped their own self-identities. The section features chapters that delve into examples such as religious conflicts during the English Reformation, clergymen's attestations to Church doctrine, aristocratic women's reading and writing habits, and the evolving ownership claims within a seventeenth-century library. Overall, it highlights the interconnectedness of individuals' lives and the narratives found within the pages of the books they interacted with. The final section is

---

[25] Acheson, *Early Modern English Marginalia*, 4.

"Modes", which investigate the concept of "mode" as a product of the intersection between a genre of writing and a technology of representation.[26] It delves into how marginalia exemplify this intersection, considering the expanse of its presence across multiple genres and technologies, such as printed books, handwriting, and conversation. Works in this section highlight how marginalia can be seen as collaborative authorship and how modes like Twitter mirror early modern marginalia in building intellectual communities. Ultimately, the section demonstrates how understanding marginalia as modes enhances our comprehension of their innovative and transformative role in communication, literature, and learning. In its entirety, *Early Modern English Marginalia* offers a comprehensive exploration of the multifaceted nature of marginalia, illuminating its significance as a dynamic lens for understanding the intricate interplay between material culture, self-expression, and modes of communication in the early modern era.

Given these understandings of marginalia, through both their presence as historical records and their physicality, the potential of engaging with these elements via digital representation suggests the richness that could be uncovered by exploring them with methods which embrace this format. The multifaceted nature of marginalia becomes even more so when a digital layer is added, opening doors to innovative modes of analysis and interpretation through data-driven investigations into patterns that might otherwise remain hidden.

---

[26] Acheson, *Early Modern English Marginalia*, 9.

Presence in Physical and Digital Archives

How is Marginalia defined?

Before discussing the presence of marginalia within the archive, it is important to understand how marginalia itself is defined. In the broadest sense of the word, marginalia encompasses anything left in the peripherals of a text, with the word deriving from the Latin *margō* ("border, edge") which evolved into the Medieval Latin neuter plural of *marginālis* ("on the periphery").[27] However, those who study marginalia typically seek to shape the meaning of it further, within the context of their own and others' research. In her 1994 analysis of annotations present in copies of Caxton's *Royal Book*, Elaine E. Whitaker proposed that marginalia tended to fall into the following three categories:

I. Editing
    a. Censorship
    b. Affirmation
II. Interaction
    a. Devotional Use
    b. Social Critique
III. Avoidance
    a. Doodling
    b. Daydreaming

Providing further context to these categories, she noted that readers would edit their texts by covering sections (A) or emphasising sections with a variety of both standard and idiosyncratic marks (B). Whitaker defines interaction with the text by the reader accepting a passage and noting how it applies to their own lives (A), or

---

[27] Oxford English Dictionary, "Marginalia, n., Etymology" (Oxford University Press, 2023), https://doi.org/10.1093/OED/7050641376.

appropriating it as a critique of someone or something else (B). She then outlines avoidance as being a subversive act, in which the reader used their text for something like the practice of penmanship (A) or recording thoughts that are not relevant to the text (B).[28] In a similar vein, Carl James Grindley in his analysis of late medieval and early modern copies of *Piers Plowman* put forward the following classifications for printed and written marginalia in texts:

> TYPE I, which comprises marginalia that are without any identifiable context;
> TYPE II, which comprises marginalia that exist within a context associated with that of the manuscript itself; and
> TYPE III, which comprises marginalia directly associated with the various texts that the manuscript contains.[29]

Grindley greatly expands on these classifications with specific sub-types assigned to each class. Type I includes marginalia such as doodles or sample texts– "short works, in either poetry or prose, which were added in an unplanned if not haphazard manner to a non-related existing text"– similar to Whitaker's category of "Avoidance".[30] Type II marginalia encompasses the space between marginalia unrelated to the text it uses as its foundation and that which reveals the "active" reader. While not providing explicit commentary on the text at hand, this form of marginalia might offer introductory materials such as brief descriptive notes identifying the main theme or subject of a work,

---

[28] Whitaker, "A Collaboration of Readers," 236.

[29] Carl James Grindley, "Reading Piers Plowman C-Text Annotations: Notes Toward the Classification of Printed and Written Marginalia in Texts from the British Isles 1300-1641," in *The Medieval Professional Reader at Work: Evidence from Manuscripts of Chaucer, Langland, Kempe, and Gower*, ed. Kathryn Kerby-Fulton and Maidie Hilmo (Victoria, BC: English Literary Studies, 2001), 77.

[30] Grindley, "Reading Piers Plowman C-Text Annotations," 79.

or marks of attribution that indicate the origin of a text. Type III marginalia tend to be the form of marginalia which historians such as Sherman and Orgel anchor their textual analysis to; this marginalia is substantial and "implies a coherent reader response to a particular text" which in turn elicits the most substantial sub-classification by Grindley. He breaks down Type III marginalia into the following:

1) Narrative Reading Aids, which denote annotations that include aids to understanding the narrative present in the text, such as citations, translations, or summation.
2) Ethical Pointers are a demonstration of ethical positionality.
3) Polemical Reponses are associated with a social or political issue in the text, either deliberating over the situation at hand or applying the situation to one that is contemporary with the commentator.
4) Literary Responses entail the reader engaging in dialogue with the text, through commenting on linguistic, humorous, ironical, allegorical, metaphorical, or other "poetic" elements of the text.
5) Graphical Responses feature systemised forms of graphic shorthand or added punctuation.

With his breakdown of classification, Grindley sought to aid the historian in answering the questions of what a particular reader was interested in, how the reader organize a text, what reactions readers had to particular passages, and if the annotations followed any general themes.[31] In the three categories which Brayman Hackel outlines in her survey of English readers, she seemingly only considers Grindley's Type III classification as being "marginalia":

> Early modern readers' handwritten marks in books generally fall into three classes, each of which exposes a set of attitudes about books and reading. Marks of active reading (deictics, underlining, summaries, cross-references, queries), *to which I refer loosely as marginalia*, suggest that the book is to be engaged, digested, and re-read. Marks of ownership (signatures, shelf marks, proprietary verses) distinguish a book as a physical object, to be protected, catalogued, inventoried, and valued. Marks of recording (debts, marriages,

[31] Grindley, "Reading Piers Plowman C-Text Annotations," 81.

> births, accounts) seem to reside somewhere in between: like ownership marks, they suggest that the book has physical value; like readers' marks, they convey that the book is a site of information. For each of these three kinds of notes, the book takes on a different role: as intellectual process, as valued object, and as available paper.[32]

For Brayman Hackel, "active" reading and evidence of intellectual process appear to be prerequisites to a mark being classified as "marginalia". Unlike previous systems of classification and as Orgel points out, Brayman Hackel also does not leave space in her classifications for the ubiquitous "irrelevant markings" despite discussing these findings in her analysis of copies of *Arcadia*[33]:

> Fragments of verse, lists of clothing, enigmatic phrases, incomplete calculations, sassy records of ownership: some of these traces merely puzzle. Drawings and doodlings in other copies hint at other associations or preoccupations: a shield painted in watercolors, impish faces peering out from the margin, geometric figures on a flyleaf, a mother and child on a blank sheet. Pens are not the only objects that have left impressions in these books; pressed flowers survive in two volumes, and the rust outlines of pairs of scissors betray the forgetfulness of the binders, presumably, of two other copies. But other marks do fall into larger patterns, joining hands across several volumes. Fifty-six percent of the books carry marginalia or scribbling on flyleaves, most commonly in the form of penmanship practice, emendations, underlinings, and finding notes.[34]

What is and is not classified as marginalia outlines a framework– both one theorised to be that of the composers, and one that the historian may follow when analysing their marks. These marks that Brayman Hackel excluded from her framework of marginalia have been conceptualized by other scholars as "graffiti". Jason Scott-Warren draws upon Fleming's work in considering the material affordance of that which

---

[32] Brayman Hackel, *Reading Material in Early Modern England*, 138.

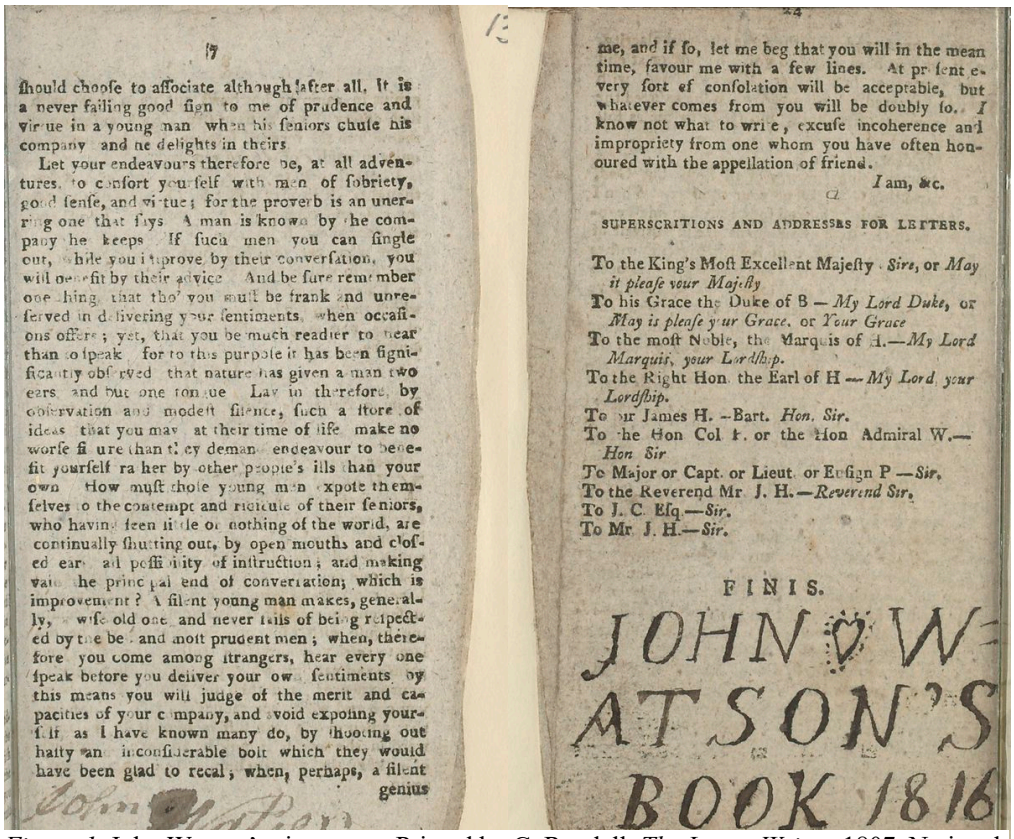
[33] Orgel, *The Reader in the Book*, 4.

[34] Brayman Hackel, *Reading Material in Early Modern England*, 159.

has been marked– these impromptu inscriptions are ones which emphasize the availability and visibility of the writing surface. Fundamentally, "graffiti" is evocative of the person who created it, the place they put it, and the documentation of a relationship between them.[35] One of the most common forms of graffiti found in early modern books is "sassy records of ownership" as Brayman Hackel calls it, the recording of names, present not in a conventional sense of simply marking ownership of the text at hand, but strewn throughout the text, written repeatedly, in a way similar to how a graffiti artist may tag buildings throughout their city. For example, in a copy of the chapbook *The Letter Writer* held in the National Library of Scotland's archive, annotator John Watson appeared to have used the pages to practice his signature, marking a number of pages with his name in both cursive and print (see Figure 1).

---

[35] Jason Scott-Warren, "Reading Graffiti in the Early Modern Book," *The Huntington Library Quarterly*, 2010, 366, https://www.proquest.com/docview/763492186/abstract/832093B897F1447APQ/1.


17

should choose to associate although after all, it is a never failing good sign to me of prudence and virtue in a young man when his seniors chuse his company and he delights in theirs.

Let your endeavours therefore be, at all adventures, to comfort yourself with men of sobriety, good sense, and virtue; for the proverb is an unerring one that says A man is known by the company he keeps. If such men you can single out, while you improve by their conversation, you will benefit by their advice. And be sure remember one thing, that tho' you must be frank and unreserved in delivering your sentiments, when occasions offers; yet, that you be much readier to hear than to speak; for to this purpose it has been significantly observed that nature has given a man two ears and but one tongue. Lay in therefore, by observation and modest silence, such a store of ideas that you may at their time of life make no worse figure than they demand endeavour to benefit yourself rather by other people's ills than your own. How must those young men expose themselves to the contempt and ridicule of their seniors, who having seen little or nothing of the world, are continually shutting out, by open mouths and closed ears all possibility of instruction; and making vain the principal end of conversation; which is improvement? A silent young man makes, generally, a wise old one and never fails of being respected by the best and most prudent men; when, therefore you come among strangers, hear every one speak before you deliver your own sentiments by this means you will judge of the merit and capacities of your company, and avoid exposing yourself as I have known many do, by shooting out hasty and inconsiderable bolt which they would have been glad to recal; when, perhaps, a silent genius

me, and if so, let me beg that you will in the mean time, favour me with a few lines. At present every sort of consolation will be acceptable, but whatever comes from you will be doubly so. I know not what to write, excuse incoherence and impropriety from one whom you have often honoured with the appellation of friend.

I am, &c.

SUPERSCRITIONS AND ADDRESSES FOR LETTERS.

To the King's Most Excellent Majesty.—*Sire*, or *May it please your Majesty*
To his Grace the Duke of B—*My Lord Duke*, or *May is please your Grace*. or *Your Grace*
To the most Noble, the Marquis of H.—*My Lord Marquis, your Lordship.*
To the Right Hon. the Earl of H—*My Lord, your Lordship.*
To Sir James H.—Bart. *Hon. Sir.*
To the Hon Col H. or the Hon Admiral W.—*Hon Sir*
To Major or Capt. or Lieut. or Ensign P—*Sir.*
To the Reverend Mr J. H.—*Reverend Sir.*
To J. C. Esq.—*Sir.*
To Mr J. H.—*Sir.*

FINIS.

JOHN W-
ATSON'S
BOOK 1816

Figure 1. John Watson's signatures. Printed by C. Randall, *The Letter-Writer*, 1807, National Library of Scotland, http://digital.nls.uk/104186662.

Sherman expands on the value of these markings well; in studying the annotated texts found in the Huntington Library, he noted how a majority of the annotations had no obvious connection with the text they accompanied, yet they "nonetheless testified to the place of that book in the reader's social life, family history, professional practices, political commitments, and devotional rituals."[36]

The archives which I draw from for this project's case study will be expanded upon much more extensively in a later section, yet at this moment it is valuable to consider how marginalia is represented in them; what is counted as marginalia in the

[36] Sherman, *Used Books*, xiii.

context of digital collections? The Archaeology of Reading (AoR) collection existed as a project to digitize and compile select annotated books from the libraries of prolific early modern readers John Dee and Gabriel Harvey, and to then transcribe, and occasionally translate, the marginalia on pages it is present so that it can be easily read alongside the printed text and archival item's metadata.[37] In studying these transcriptions, AoR includes any category of textual marginalia of which there is plenty due to the interactive approach Dee and Harvey took to reading. The more elaborate graphical responses such as manicules and florilegia are also made note of, however symbols such as small crosses or underlinings are excluded despite (or perhaps because of) their abundance. These symbols are often excluded from broad analysis of marginalia despite being a constant presence found almost exclusively in the margins of text– they are marginalia in the truest sense of the word. Conversely, the Early Modern Annotated Books collection from UCLA's William Andrews Clark Memorial Library only identified that the book, or, collection of pages, had marginalia within them, but did not indicate which pages and by extension, offer any transcription of the marginalia. What is defined as marginalia here is entirely up to the reader of their archive.

Physical and Digital Presence

In considering the archive, marginalia have always been present yet their presence has always been contentious. Early modern readers were taught to read with a writing tool in hand; they were not just encouraged to annotate their books, but taught how to do

---

[37] "Archaeology of Reading," September 2014, https://archaeologyofreading.org/.

so in their schooling.[38] Marginalia was a way for readers to connect with their previous knowledge, literature, and each other as the book circulated, yet as the concern exists at present, there was concern even during the early modern period about, as Sherman phrases it, “dirty books”. Books with marginalia were considered marred, being no longer in the pristine condition that they were in fresh off the printing press, which was a largely desired state for collectors from the early modern period into the 20th century.[39] The notable exception to this rule being annotations made by famous figures such as William Shakespeare or Francis Bacon, which added value to the books, that is, if they were identified before being “cleaned”. Due to the desire for virgin books, annotated books were often “restored” in a way that removed the marginalia present through methods such as the bleaching of page edges or by cutting out the margins of these books then rebinding them. This is all to say, the marginalia present in physical archives around the world is either skewed to portraying to the reading and writing practices of a select educated and famed class of society, preserved due to the perceived significance of the books they owned, or accidental by-products of archiving used books. This historical lack of intentional curation of marginalia has made the exploration of them a challenge. There are no single physical archives that gather marginalia en masse to be studied, they are more often discovered organically when studying the content of early modern books, or by paging through large collections of early modern books with the intention of searching

---

[38] Sherman, *Used Books*, 4.

[39] Sherman, *Used Books*, 159.

for marginalia specifically. Since marginalia is largely left uncatalogued, it remains difficult to find even during the advent of the digital archive.

In 2008, reflecting on the laborious process he undertook when producing his expansive text on early modern marginalia, Sherman wrote in the afterword of *Used Books*:

> Databases and facsimiles of the sort described above are primarily concerned with giving us access to accurate and attractive informational content and with helping us to make our way around it (a goal generally known, in the computer and information sciences, as "usability"). Their emphasis on "interactivity" notwithstanding, they have not yet imagined us doing much with or to books beyond turning their pages and have not yet found ways to preserve our marks—much less to improve them or to educate us about the markings of those who turned pages before us.[40]

As of the year I am writing this, 2023, despite many technological advances in the realm of the dead hand of the pdf and skeuomorphic design, Sherman's statement holds largely true within the bounds of digitized cultural heritage collections. AoR is the only project I discovered in the process of research that places marginalia in the context of "those who turned pages before us," yet its functionality does not go beyond turning from page to page, and simple manipulations such as zooming in and out. The one project I did come across in which 3D models were constructed from the pages of manuscripts found at Lichfield Cathedral had marginalia make an accidental appearance.[41] As a whole, marginalia is still not easily discoverable even with digitised archives outside of projects dedicated to identifying marginalia, which are also few. Partly, this is due to the way

---

[40] Sherman, *Used Books*, 182.

[41] Noah Adler and Justin Hall, "Matt 28:19 - 28:20, Pg 141," *Manuscripts of Lichfiled Cathedral*, accessed August 21, 2023, https://lichfield.ou.edu/file/14428.

digital technologies have evolved to prioritize text (not margins), and also partly due to metadata standards which shape digital archival items failing to adapt their form to the unique affordances that digitisation offers; metadata standards for digitisation focus on documenting details about the physical object and were developed from the same cataloguing standards that excluded marginalia originally. AoR explains this problem as one they faced during the development of their project. Upon digitizing the books of Dee and Harvey, AoR then needed to select a standard to use when creating the digital documents. Initially, they looked at the Text Encoding Initiative (TEI) standard, formed by a consortium which collectively develops and maintains a standard for the representation of texts in digital form. The TEI is largely considered a high standard of documentation for representing digitised texts, yet upon further investigation AoR found that TEI left no option for recording annotations featured on a text, thus they ultimately had to create their own bespoke digitisation schema for their project which centered around this form of text.

As I will demonstrate, there are now computational methods that could be used by archives to discover marginalia in their collections and in turn, an opportunity to automatically enhance their metadata. However, these methods require an extensive amount of "examples" of marginalia in order to begin identifying it; these examples are difficult to accumulate due to the lack of any indication of marginalia within digital archives, thus creating a cycle in which marginalia continues to be an elusive presence.

On Object Marks and Machines

Acheson's book raises a number of historical, cultural, and theoretical questions about books and their readers, many of which could be explored further through greater access to more obscure, both literally and figuratively, research materials such as marginalia. Adam Smyth's contribution to *Early Modern English Marginalia* particularly emphasises this; in his chapter, he sought to identify current traits in recent works on the history of marginalia which bring to the surface a number of questions, paradoxes or gaps, through the analysis of object traces in early modern books.[42] These traits-as-problems which Smyth uses to illustrate the value of attending to object traces can also be used to understand the value of integrating machine learning methods, particularly object detection, into the study of marginalia.

The history of reading has often revolved around individual readers, even as early modern studies broadly have shifted away from this and towards the inner circle of past scholars, and, in more recent years, to the network as the unit of cultural analysis.[43] This focus on biography has meant that studies of marginalia have tended to connect book annotations back to the individual who created them at the expense of other ways of organizing marginalia, such as by genre. Further expanding on the neglect of genre, the use of biography as the frame for analysing marginalia books has meant that scholars, preoccupied connecting the page marks to the reader's life and identity, have spent less time seeking to establish fundamental histories of the marks themselves, such as where

---

[42] Acheson, *Early Modern English Marginalia*, 63.

[43] Acheson, *Early Modern English Marginalia*, 64.

the conventions for marking books came from; in seeking to link marginalia to writing outside of the host text, Smyth posits the questions, "What category of mark or intervention are they? What is the larger group in which they belong?"[44] The tether which holds the study of marginalia to biography is at least in part due to the issues of discoverability that marginalia presents. Although more reliably discoverable than the trace object marks which Smyth features in his work, marginalia still demand time to be found whether the host being searched is physical or digital. Identifying genre or larger categories which marginalia can be viewed through or belong to require wide reading of numerous sources in order to be defined. Undoubtedly, the work of studying and defining marginalia to date has been an impressive feat, but as noted by the historians who performed this feat, it is immensely time consuming and difficult process if the marginalia one seeks to study is not already gathered, which is often the case. The automated detection of marginalia that my tool permits allows for large and diverse corpora to be evaluated for the presence of marginalia with little supervision and more efficiently than a researcher is able to sift through these works manually; following this quick process of identifying marginalia, researchers may then focus on the discovery of genre through analysis of the object detector's output.

Another trait Smyth identifies in the study of marginalia is the assignment of the reader as "active". Rather than being "passive", readers read with an idea of practical application in their world and the future, "reading as intended to give rise to something

---

[44] Acheson, *Early Modern English Marginalia*, 64.

else."[45] Smyth refers to the inverse of an active reader, an *inactive* reader, as being found present in unmarked pages, yet I propose an inactive reader could also be defined as one who dismisses the content of the book in favour of its materiality, as something that can be repurposed as a diary or catalogue of debts (cite wigmaker's pages). It is this genre of "inactive" reading that I attempt to find in the following case study using an object detection model applied to the NLS collection of early modern chapbooks printed in Scotland.[46] By microscopely analysing the marginalia of over 3000 works made for popular consumption, trends in how these booklets were used as something more than just the entertainment found in their contents can be identified.

Creation of an Application to Capture the Metadata of Big Data

Given the complicated and multifaceted ways marginalia can be read and understood, how their complex materialities require close observation and tactile engagement with the page, and how the definition of what 'counts' as marginalia can be so contested, an application designed to identify and find marginal annotations might seem a foolish endeavour. If we shift perspective for a moment, from the margins of our books to the margins of our planets, a similar *kind* of problem can be seen for archaeologists who study human settlements in space (admittedly, there is only one single such settlement at present, the International Space Station). That is to say, the problem is one of identifying the interesting elements in a collection of materials where we cannot

---

[45] Jardine and Grafton, ""Studied for Action"," 30.

[46] Scotland, "Chapbooks Printed in Scotland".

physically examine the 'real' materials. A problem at present for space archaeologists is that they study the margins of lived lives, the detritus left over from earth beyond its stratosphere, yet NASA alongside other space agencies do not permit archaeologists to become astronauts.[47] So, the archaeologists interested in how life is lived on the ISS can only work from photographs taken by those who were able to become astronauts. For the archaeologists to study these photographs, they needed a way of annotating the images so that larger patterns could be deduced. This context alongside discussions on metadata and machine learning within cultural heritage institutions are what framed the first technical component of my MRE.

My application for identifying annotations to create a training corpus is actually a further adaptation of an application I created for the International Space Station Archaeology Project (ISSAP) to support the needs of archeologists working on the project.[48] The ISSAP version of the application is used to analyze photos taken of the living quarters ISSAP received from the International Space Station. The project sought to understand how astronauts use the space of the habitation modules by tracking the small items of daily use across the station as they appeared and disappeared in photographs.[49] The archaeologists annotate photos to eventually create an 'automatic

---

[47] Rachael Blodgett, "Frequently Asked Questions NASA," Text, *NASA*, January 2018, http://www.nasa.gov/feature/frequently-asked-questions-0.

[48] ChantalMB, "ChantalMB/Issap-Image-Annotator," February 2022, https://github.com/ChantalMB/issap-image-annotator.

[49] Shawn Graham and Justin Walsh, "Recording Archaeological Data from Space," *International Space Station Archaeological Project*, February 2022, https://issarchaeology.org/how-do-you-get-from-an-astronauts-photo-to-usable-archaeological-data/.

archaeologist' that can analyze spatial patterns of material culture in the photos. The tool I created for ISSAP is a rewritten version of the more general purpose Visual Geometry Group Image Annotator developed at Oxford University; the original tool was not suitable for the project because it could not be used collaboratively and did not have a structure for the automated recording of metadata as annotations are generated.[50] In general, image annotators are used when creating applications for computer vision to create datasets for the purpose of training classification or detection based machine learning models to recognize items of interest. The researcher annotates the image, and the machine learning model learns to look for these features which have been marked as important. A side effect of this approach is that the metadata created by the researcher while producing these annotations to train the model with becomes divorced from the original metadata of the images. When considering this in relation to the study of book marginalia, it is equivalent to cutting the marginalia out of the pages and only analyzing those select segments without any thought about the document the marginalia came from.

I sought to expand on the collaboration functionality as well as the ability to import existing metadata from both the archive and directly from the images to be annotated in this new version of the application, which I named RocketAnnotator as a gesture towards its galactic origins. With this feature, new metadata can be produced alongside the context of the original object metadata common in archives, allowing for

---

[50] Abhishek Dutta and Andrew Zisserman, "The VIA Annotation Software for Images, Audio and Video," in *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19 (New York, NY, USA: Association for Computing Machinery, 2019), 2276–79, https://doi.org/10.1145/3343031.3350535.

the tool to become not only an image annotator, but also a way to reference and track the creation of training data for machine learning projects. Additionally, all metadata will be easily searchable both from within the application and through the structured output file, allowing for researchers to easily reference specific data points and for scholars to browse the data which formed the output being presented to them.

In 2016, book historian Ryan Cordell called for more robust methods to describe digital artifacts bibliographically within the context of utilizing digitised archives. Research which makes use of these digital objects often fails to account for the sources, technologies, and social realities of the objects' creation in ways that make their affordances and limitations more readily visible and available for critique.[51] Likewise, in conceptualizing digital archives as sources of data in their book, *Data Feminism*, Catherine D'Ignazio and Laura Klein continuously emphasize the necessity of further context at all stages of working with "data", from acquisition to analysis, because the context which data is situated in is seen as essential to the ultimate "framing and communication of results" formed through its use.[52] How a digital object is catalogued within the archive becomes how the object is situated when used as data, thus when bibliographic records only offer an incomplete account of the digitised object the results of research using this data are also incomplete in a way, as the researcher is not provided

---

[51] Ryan Cordell, ""Q i-Jtb the Raven": Taking Dirty OCR Seriously," *Book History* 20, no. 1 (2017): 191, https://doi.org/10.1353/bh.2017.0006; Bonnie Mak, "Archaeology of a Digitization," *Journal of the Association for Information Science and Technology* 65, no. 8 (2014): 3, https://doi.org/10.1002/asi.23061.

[52] Catherine D'Ignazio and Lauren F. Klein, *Data Feminism* (Cambridge, MA: The MIT Press, 2020), 164, https://doi.org/10.7551/mitpress/11805.001.0001.

all the information necessary to understand the object in its entirety. This line of thought signals that for any tool designed with the use of digital artifacts and collections in mind, it is vital to attempt at including ways that this lost data could be drawn out and made accessible to the user.

When archival materials are integrated into research utilizing more traditional methods of historical inquiry, the subject being analyzed tends to be singular – focused on the work of one individual or the content of one collection. This in turn makes answering questions about the affordances and limitations of their sources more manageable without extensive organization, since there is cohesion across sources. Comparatively, the large amount of data needed for machine learning methods often results in these questions being difficult to answer on a microscopic level, because of both the diversity in the data when drawing from multiple sources and the archival metadata being omitted during the process of data collection. The scale of the dataset produced makes such detailed information be perceived as unnecessary, a mode of thought which carries even into projects with humanistic foundations.

Yet these details which go into the first step of building a machine learning model are vital to understanding the influences and limitations of it; the foundations which machines learn with and from are human, meaning that they contain "human subjectivities, biases, and distortions" like all other works created by humans.[53] In order to produce meaningful output using either analog or automated research methods, such as

---

[53] Benjamin Lee, "Compounded Mediation: A Data Archaeology of the Newspaper Navigator Dataset," *Digital Humanities Quarterly* 015, no. 4 (December 2021), http://www.digitalhumanities.org/dhq/vol/15/4/000578/000578.html.

machine learning for identification, it is vital to interrogate the input that contributed to the making of the method being applied for answers regarding social, cultural, historical, institutional, and material conditions under which that input was produced, as well as about the identities of the people who created it.[54]

The annotation editor and metadata viewer which occupies most of RocketAnnotator's bottom pane was created specifically to address these issues of decontextualized data in large scale datasets. At the surface level, it is designed to appear similar to a spread sheet such as those found in Excel, so it is intuitive to the user understanding what the section of the application is for and how it is used. In the first tab, "Annotations", there are five descriptive columns present by default: the annotation's unique ID, the date and time the annotation is created, who the annotation is created by, the broader category the annotations fall into, and what specifically the annotation is. There is a "+" symbol at the end of the column headers that allows the user to extend this metadata through adding their own columns specific to the project. A row is added to this table each time an annotation is drawn on an image, and likewise, deleting an annotation on the image canvas deletes the corresponding row, facilitating a direct connection between the image and the metadata being generated. The second tab, "Metadata", displays data associated with the image being annotated– this can be metadata from the digital archive which the image was obtained from if the archive chooses to tag their images with this information or the user does so in the process of collecting the images,

---

[54] D'Ignazio and Klein, *Data Feminism*, 152.

as well as any additional EXIF data, the metadata embedded within digital images, which can be extracted.

Understanding Metadata through its Usage

Cordell encourages us to think of items found within digital archives as not simply a transparent surrogate for a corresponding physical object, but instead as a "new edition" in the full bibliographic sense of the word; while it "departs more and more from the form impressed upon it by its original author," it nonetheless "exerts, through its imperfections as much as through its perfections, its own influence upon its surroundings."[55] When it comes to cultural heritage collections, the digitised item is often described in metadata as if it were the original item picture rather than a new version; in museums, replicas of deteriorated artefacts are marked as such, yet digitised objects are often treated as if they are exact substitutes for the physical. As Adam Crymble demonstrates in his history of mass digitization, the digitisation of primary sources was to a great extent driven by the desire to democratize primary sources for education and research purposes; in the beginning, digitised sources *were* explicitly intended to be surrogates for the original.[56] By and large, digitised sources have been used as such, and so this form of metadata has been considered suitable for its audience. Yet in the age of big data and machine learning, the digital archive's audience has shifted from solely human consumption to machine consumption as well. Archival metadata and what that

---

[55] Cordell, "'"Q i-Jtb the Raven"'", with quote from W. W. Greg.

[56] Adam Crymble, *Technology and the Historian: Transformations in the Digital Age* (Champaign, IL: University of Illinois Press, 2021), 68, https://doi.org/10.5406/j.ctv1k03s73.

entails must be expanded to fit this use. Metadata, the data which describes data, is what holds data accountable.

In the context of machine learning, what has been perceived as valuable is the data that will be used to train a model, and any data surrounding that data is largely ignored or discarded after it is finished its use in create the training data. A model uses an algorithm to make sense of the data given to it, and produce some form of task or output, such as classifying images or generating a paragraph of text. In order for an algorithm to adequately "learn" to do something, it needs an extensive number of examples; for example, the Common Objects in Context (COCO) dataset is a popular dataset used for training object detection models, and it contains 1.5 million examples of objects in photos which fall into one of 80 categories.[57] The development of these massive datasets nearly always involves ingesting massive amounts of data from convenient or easily-scraped Internet sources such as Twitter or Flickr under the assumption that this will inherently result in diverse content, therefore metadata serves little purpose since the data was not created by nor possible to be revised in its entirety by a person.[58] The datasets which do offer metadata associated with the items rarely offer it in an accessible manner, with metadata for the mass amounts of content being stored in obscure file formats or in large multipart archives.[59] This belief in the unimportance of metadata has resulted in

---

[57] Tsung-Yi Lin et al., "Microsoft COCO: Common Objects in Context" (arXiv, February 2015), https://doi.org/10.48550/arXiv.1405.0312.

[58] Bender et al., "On the Dangers of Stochastic Parrots," 613.

[59] Andy Baio, "Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator," *Waxy.org*, August 2022,

researchers lacking an understanding of the training data being used to train their models, which has led to multiple instances of machines learning to replicate the harmful views their data possess. A recent example at the time of writing this would be the Stable Diffusion text-to-image generation model, which was trained on billions of image-text pairs scraped from across the internet.[60] Claims from both casual users and formal investigation of the model have found that Stable Diffusion may unexpectedly generate inappropriate or disturbing images, as well as otherwise offensive content; for example, images generated using the statement "Japanese body" yielded almost exclusively inappropriate material, with 90% showing explicit nudity.[61] Closer attention paid to the metadata of the training data could have mitigated undesirable outcomes and identified patterns of discrimination before they were fed to the model and reproduced.

In recent years, there has been movement within the field of computer science towards critical analysis of how datasets are constructed, composed, and used. Primarily, these efforts have been directed toward standardizing the documentation of datasets through 'datasheets', overviews attached to datasets which communicate the content of a

---

https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/.

[60] Robin Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models," 2021, https://arxiv.org/abs/2112.10752.

[61] Patrick Schramowski et al., "Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models," in *Proceedings of the 22nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Vancouver, BC: arXiv, 2023), 3, https://doi.org/10.48550/arXiv.2211.05105.

dataset in a way that prioritizes transparency and accountability.[62] Within this conversation, there has been encouragement to draw upon the existing language and procedures for managing sociocultural data within libraries and archives. In *Lessons from Archives*, scholars Eun Seo Jo and Timnit Gebru argue that archives, as "a form of large-scale, collective human record-keeping", can aid in addressing the questions of power imbalance, privacy, and other ethical concerns that datasheets leave unaddressed through interventionist data collection strategies to address biases and ensure fair representation.[63] They indicate a number of ways they believe practices which emerge from archival studies would enhance the practice of machine learning; firstly, that archives begin with focused, institutional mission statements that outline a commitment to "collecting the cultural remains of certain concepts, topics, or demographic groups" which guides their data collection process, as well as curators who are responsible for weighing the risks and benefits of gathering different types of data in relation to an archive's objectives and have developed theoretical frameworks for appraising collected data.[64] Gebru and Jo encourage the machine learning community to approach data collection and appraisal by at least starting with a statement of commitment rather than starting with datasets by availability to ensure equitable targets during the construction of datasets. This echoes

---

[62] Timnit Gebru et al., "Datasheets for Datasets," December 2021, 10, https://doi.org/10.48550/arXiv.1803.09010.

[63] Eun Seo Jo and Timnit Gebru, "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20 (New York, NY, USA: Association for Computing Machinery, 2020), 2, https://doi.org/10.1145/3351095.3372829.

[64] Jo and Gebru, "Lessons from Archives," 5.

D'ignazio and Klein earlier conceptual call for data scientists to proceed with awareness of context and an analysis of power in the collection environment to determine whose interests are being served by being counted in the dataset, and who runs the risk of being harmed.[65] Additionally, archives often have codes of conduct or ethics and a professional framework for enforcing them alongside developed detailed standards for data description, ensuring ethical practices in data collection by helping ensure transparency and accountability; these multi-faceted forms of review and record-keeping are unheard of in machine learning data collection.[66] Lastly, archival sciences have promoted collective efforts to address issues of representation, inclusivity, and power imbalance; for example, community-based activism has been used to ensure that various cultures are represented in the manner in which they would like to be seen.[67] Machine learning researchers can draw from these efforts towards participatory archives to ensure diverse and inclusive datasets.

As *Lessons from Archives* highlights, the issues of historical power structures and how they may be dismantled have long been discussed within archival studies. Yet what *Lessons from Archives* does not discuss is the digital turn within the archive itself, how the archive, through embracing a digital form, has moved from being a collective of

---

[65] D'Ignazio and Klein, *Data Feminism*, 111.

[66] Jo and Gebru, "Lessons from Archives," 7.

[67] Jo and Gebru, "Lessons from Archives," 5.

human recordkeeping to a collection of data to be made sense of and mined.[68] When viewing collections themselves as data, archival data is seen as beneficial for the existing metadata associated with or describing each archival item. Unlike a blog post where metadata about it needs to be constructed by identifying and compiling available information from the web page, items in a collection have this descriptive information already curated and compiled. Yet a significant issue comes from this perception of digital archives broadly as complete in their current state. Despite institutional mission statements, codes of ethics, and community contributions, at an individual item level, the metadata is still the same as that in catalogs which have long represented groups of people in problematic ways. What has changed is the new methodology being used to promote the use and reuse of these descriptions and collections. As librarian Sophie Ziegler writes in their article *Open Data in Cultural Heritage Institutions: Can We Be Better Than Data Brokers?*, "The collections as data framework in cultural institutions carries with it the possibility for our descriptions of people to be shared, combined with other data, and used to negatively affect groups."[69] When framing the archive as data, there is a risk that the archival holdings and their descriptions will look objective and natural, and the work of archivists and others to show how archival collections are never neutral and natural will be obscured; Devon Mordell encourages "active participation and

---

[68] Michael Moss, David Thomas, and Tim Gollins, "The Reconfiguration of the Archive as Data to Be Mined," *Archivaria*, November 2018, 131, https://archivaria.ca/index.php/archivaria/article/view/13646.

[69] S. L. Ziegler, "Open Data in Cultural Heritage Institutions: Can We Be Better Than Data Brokers?" *Digital Humanities Quarterly* 014, no. 2 (June 2020), http://www.digitalhumanities.org/dhq/vol/14/2/000462/000462.html.

critical discourse" around the tools and practices to ensure that new technologies reinscribe this false sense of neutrality.[70] One simple yet significant action that works toward this goal is incorporating data process and provenance into the standardized documentation practices for collections. During his time as Humanities Data Curator at the University of California Santa Barbara, Thomas Padilla emphasized the concept of *legibility* within metadata, that to make collections as data usable, the processes behind their establishment must be transparent and documented. In the context of libraries, Padilla indicates that:

> Libraries do not often provide access to the scripts that generate collection derivatives, access to processes for cleaning or subsetting data, access to custom schema that have been used, indications of how representative digital holdings are relative to overall holdings, nor is the quality of data typically indicated. Libraries do not typically expose why some collections have been made available and others have not. Libraries do not typically identify the library staff personally responsible for modifying, describing, and creating collections – a dimension of provenance that must be accessed in order to determine data ability to support a research claim.[71]

These same claims can be applied to archival items. Without this information, the user's ability to comprehend and thus utilize a collection as data is hindered or even made impossible through the elusive gaps which are left in the collection that would then be transferred to any project that makes use of it. The data is left vulnerable to misuse when

---

[70] Ziegler, "Open Data in Cultural Heritage Institutions"; Devon Mordell, "Critical Questions for Archives as (Big) Data," *Archivaria* 87 (2019): 156, https://proxy.library.carleton.ca/login?qurl=https%3A%2F%2Fwww.proquest.com%2Fscholarly-journals%2Fcritical-questions-archives-as-big-data%2Fdocview%2F2518871266%2Fse-2%3Faccountid%3D9894.

[71] Thomas Padilla, "On a Collections as Data Imperative," 2017, 3, https://escholarship.org/uc/item/9881c8sv.

not fortified through comprehensive metadata. The potential of collections as data hinges on integrity validated through expanded documentation practice.

The annotation editor and metadata viewer within my application seeks to address the digital archive in the state it is at present. The level at which data provenance is addressed varies widely from institution to institution, thus there are features built into the application which seek to close some of the gaps surrounding the digital origins of objects being annotated. The metadata viewer shows both how the image was contextualised within the archive it was extracted from, and since many archives have not yet begun to include information about the entry as a digital object, the application includes the automatic extraction of EXIF data to expand the predefined archival metadata. EXIF data can provide details on camera settings including make and model, date, and time the image was captured, geographic information regarding where an image was taken, photography settings such as white balance or flash usage, and information on what software was used to process or edit the image. Essentially, a potentially detailed history of how an image was captured and processed when this information might not otherwise be present. Being able to view both the archival and digital metadata in the process of annotation ultimately aides in circumventing the decontextualized access and consumption which occurs during the process of annotating data for computational research.[72]

In light of discussion over digitised archival objects being a new edition in the lineage of an item, the annotation editor expands on this mode of thought and encourages

---

[72] Milligan, “We Are All Digital Now,” 617.

the annotator to view their annotations in the same way. Each annotation visually segments a portion of the image from its surroundings, marking it as something significant which is important enough to be highlighted and thus it should be documented in a way that is similar to other digital objects. One way humanists can distinguish themselves in the process of creating datasets with the end goal of machine learning is through the addition of explanations about decisions that we make while creating data.[73] While the ability to add their own columns in this tab encourage the user to create structured metadata for their annotations, even without adding additional columns, the user still must record who created the annotation and basic descriptive information about the contents of the annotation, capturing key metadata which holds the creator of it accountable in the process of creation for each annotation. Treating annotations as new digital objects both enhances familiarity with the training data and constructs a more robust log of the training data with more findable items should an issue arise during the process of or after training a model.[74]

Case Study: Finding Early Modern Marginalia

Understanding Big Data

When regarding collections as data, I have discussed the challenges faced relating to metadata, but what about the issues encountered when using collections as data in their entirety? Machine learning models require large-scale data for training, fine-tuning, and

---

[73] Ziegler, "Open Data in Cultural Heritage Institutions".

[74] For discussion on limitations of the application at present, see the application's GitHub repository: https://github.com/ChantalMB/MRE-RocketAnno

evaluation. In the context of cultural heritage institutions, creating such large-scale datasets carries legal and logistical implications, not to mention the further challenge of having trained historians conversant in the techniques.[75] These compounding factors have so far slowed the potential of machine learning models for historical inquiry. Existing digitised cultural heritage collections can help bridge this gap; however, this requires that the institutions which hold this data be active participants in this shift to making their digitised collections open and accessible to use in ways other than simply viewing.

Although cultural heritage institutions are increasingly digitizing their collections and making them available through online portals for public consumption and discovery, the usability of their collections as data is rarely straightforward. Application Programming Interfaces (APIs), virtual bridges which enable applications to send and exchange data or functionality, are becoming more standardized and prevalent within digital collections. Many APIs lack comprehensive documentation on their usage beyond internal data retrieval, although large museums with more resources to allocate to digitisation are working towards developing relevant documentation or usage guides. The Victoria and Albert Museum (V&A) Collections API was one of the first to be made open to researchers, with development beginning in 2009; it is an excellent model of what documentation should look like, with detailed written guides alongside sample code

---

[75] Clemens Neudecker, "Cultural Heritage as Data: Digital Curation and Artificial Intelligence in Libraries," in *Proceedings of the Third Conference on Digital Curation Technologies (Qurator 2022), Berlin, Germany, Sept. 19th-23rd, 2022*, ed. Adrian Paschke et al., vol. 3234, CEUR Workshop Proceedings (CEUR-WS.org, 2022), 5, https://ceur-ws.org/Vol-3234/paper2.pdf.

and additional resources for their anticipated users.[76] Yet even the V&A is not exempt from one of the pitfalls which APIs of large collections face. They tend to end up being prescriptive and restrictive, particularly when it comes to allowing a user to download the entirety of the available digitised collections. The V&A API is not recommended for bulk data export, and other institutions such as the Getty do not provide a way to get a list of all of the objects or a way to download all the data in the collection at all.[77] While APIs offer a more formal way of accessing collections as data, many of those who seek to use collections as data find "simple download dumps" more useful to quickly explore what is a collection offers. Downloading a file containing the data provides direct access without the barrier of having to learn the intricacies of an API.[78] Yet even when cultural heritage institutions do offer their data as downloadable content, it often ends up being in complex schema formats like METS, MODS, and ALTO XML. While these formats are standard in the library and archival domains, they pose a barrier for use in contexts of data analysis where formats such as CSV and JSON files are preferred due to the availability of programming libraries that can easily process them.

As discussed in the context of metadata, the lack of attention paid to the curation of training data in machine learning has been a significant topic of discussion as of recent, with research emerging from the field of computer science on recommendations

---

[76] Victoria and Albert Museum, "Victoria and Albert Museum Collections Data," 2021, https://collections.vam.ac.uk/.

[77] "Getty API Documentation," *Getty*, accessed August 18, 2023, https://data.getty.edu/museum/collection/docs/#attribution.

[78] Neudecker, "Cultural Heritage as Data," 3.

for ethically sound and transparent standards for publishing datasets alongside calls for more "accountable" curation, as is perceived to be practiced in cultural heritage institutions. These discussions have fed into reflective research within archival studies on what ethical issues may arise when using cultural heritage data for the purpose of machine learning. Collections, in both digital and analogue form, are not just sources of history but also "its subjects, sites with histories and politics of their own."[79] Without critical reflection on how collections have been curated in past and present, or what has and *has not* been digitised, there is risk of models trained on cultural heritage collections to exclude entire histories if careful attention is not paid to the composition of a collection. Conversely, there is also risk of including items in the data that are by-products of colonial and exploitative histories of the archive, featuring vulnerable people who may not have chosen to have their presence be displayed in this context. Even with good intentions, projects using cultural heritage data "risk kitschifying or exploiting those represented in the digitised collections in question."[80]

A recent demonstration of contested subjects and vulnerable histories within digital archives is the Zealy Daguerreotypes, a series of photographs taken by Joseph T. Zealy featuring enslaved men and women in various states of undress commissioned by naturalist Louis Agassiz in 1850 as part of his effort to document physical evidence of polygenism, the theory that different racial groups do not share a common biological

---

[79] Elizabeth Yale, "The History of Archives: The State of the Discipline," *Book History* 18 (2015): 332, https://doi.org/10.1353/bh.2015.0007.

[80] Benjamin Charles Germain Lee, "The 'Collections as ML Data' Checklist for Machine Learning and Cultural Heritage," *Journal of the Association for Information Science and Technology* n/a, no. n/a (May 2023): 2, https://doi.org/10.1002/asi.24765.

origin.[81] These daguerreotypes were uncovered in the attic of Harvard University's Peabody Museum of Archaeology and Ethnology by an archivist in 1977, and are now featured in the Peabody Museum's online collections, being, in fact, the first of the search results for "daguerreotype".[82] Tamara Lanier, a descendant of two of Agassiz's "subjects", sued Harvard in 2019 for unlawfully possessing and profiting from the image of her ancestors, who as slaves could not have consented to these photos being taken and their further usage.[83] Despite the contention surrounding these images, when viewing their entries within the digital archive there is no indication of the history associated with these photos in the metadata, no context given about the enslaved subjects or the fact the photographed individuals were enslaved, nor the photos' purpose as "proof" of polygenesis by Agassiz. There is only a short disclaimer about historical language, which appears to be present on all items in the Peabody Museum's online collections. These individuals are not safe from being commodified again; as images in the public domain, they could easily end up in scraped into a dataset and used for the purpose of machine learning. For instance, following the controversy surrounding Stable Diffusion's questionable outputs, technologist Andy Baio and Datasette creator Simon Willison produced a searchable data browser for a sample of approximately 12 million images

---

[81] Bonde Thylstrup et al., *Uncertain Archives*, 75.

[82] "Results – Search Objects – eMuseum," *Peabody Museum of Archaeology & Ethnology*, accessed August 21, 2023, https://collections.peabody.harvard.edu/search/daguerreotype/objects.

[83] Bonde Thylstrup et al., *Uncertain Archives*, 75.

used in the training of Stable Diffusion.[84] This is about 2% of the 600 million images used to train the most recent update of the model, and only 0.5% of the 2.3 billion images that it was first trained on, yet when searching the word “daguerreotype”, a picture of a woman with her enslaved child servant feature more than once.[85]

In both an effort to keep this case study accountable as a project which uses cultural heritage collections as data for machine learning, and as a demonstration of how standardized guidelines can be implemented and beneficial to research, the remainder of this section will follow the “Collections as ML Data” checklist for machine learning and cultural heritage recently developed by Benjamin Lee.[86] Observing the growing trend in the field of machine learning to develop guidelines, checklists, and best practices for researchers and practitioners involved in creating datasets, training models, and implementing machine learning systems, Lee proposes the creation of his “Collections as ML Data” checklist. This checklist is intended to help researchers working on machine learning projects involving cultural heritage collections through addressing potential challenges such as misrepresentation, oversimplification of digitisation nuances affecting model performance, unnecessary use of machine learning, lack of sustainability planning, and privacy violations. By incorporating this checklist into their projects, researchers can engage more thoughtfully with these challenges and enhance their impact on the field of

---

[84] Baio, “Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion’s Image Generator”.

[85] “Laion-Aesthetic-6pls: Images 1582553,” *LAION-Aesthetics V2 6+*, accessed August 21, 2023, http://laion-aesthetic.datasette.io/laion-aesthetic-6pls/images/1582553.

[86] Lee, “The ‘Collections as ML Data’ Checklist for Machine Learning and Cultural Heritage”.

digital cultural heritage.[87] Lee's article draws from both machine learning and cultural heritage research, creating an interdisciplinary tool for responsible data practices.[88]

The Cultural Heritage Collection as Data

There is a distinction between the cultural heritage collection being studies, and the training dataset from which a machine learning model is created reflecting that collection.[89] In this case study, I use machine learning to train an object detection model to identify marginalia on pages of the National Library of Scotland's (NLS) collection of chapbooks; such a trained model could be used further on other early modern texts emerging from a similar context as the chapbooks. In order to create the training data necessary to teach the model what marginalia looks like, I drew from three cultural heritage collections which contained annotated texts created largely during the early modern period, with the oldest printed in the late 15th century and the occasional chapbook being from the 19th century. Pages from the latter date range were not used in the training data, but due to the parsimonious and cyclical nature of chapbooks their appearance differed only slightly if at all from chapbooks printed in earlier centuries.

---

[87] Lee, "The 'Collections as ML Data' Checklist for Machine Learning and Cultural Heritage," 2.

[88] The checklist in its original form is attached as an appendix to this project. Due to the nature of this project being an MRE, I did not answer all questions relating to organizational concerns since I am producing this MRE outside of a corporate or workplace setting.

[89] Lee, "The 'Collections as ML Data' Checklist for Machine Learning and Cultural Heritage," 19.

My interest in these texts was spurred by a previous project that was meant to identify the woodcut illustrations scattered within them; instead, alongside illustrations the detector would also pick up marginalia. Since the ultimate goal for this case study is to intentionally identify marginalia in this collection of chapbooks, I chose to incorporate these pages I had made note of that were outputted by the illustration detector into the training data for both diversity and to provide domain specific examples in hopes of improving the trained model's performance. The second digital archive drawn upon is *Archaeology of Reading* project which was a collaborative effort between the Sheridan Libraries, Centre for Editing Lives and Letters, and Princeton University library that resulted in a corpus of thirty-six fully digitised versions of early printed books annotated by "two of the most dedicated readers of the early modern period: John Dee and Gabriel Harvey".[90] Considering himself a scholar of science, the marginalia composed by Dee feature tables, charts, and diagrams to make sense of what he read. In contrast, Harvey was a humanist and approached his selection of reading as such with passages highlighted through underlining and notes connecting the text to other works read. Finally, to further increase the variety of marginalia that the model will learn from, I drew upon a collection of Early Modern annotated books compiled by the William Andrews Clark Memorial Library at the University of California, Los Angeles which spans from the 15th century up until the early 19th century as the early modern period came to a definitive close. The marginalia in this collection are the most diverse of these

[90] "Archaeology of Reading".

sources, containing the scrawl of not only scholars, but printers and other members of the community as well.

NLS Chapbooks in the Data Foundry

To ensure a complete understanding of what each collection offers researchers as both a subject to be studied and as data, we will interrogate each more deeply than just description, beginning with the NLS's chapbooks. NLS Chapbook collection finds its origins within the Lauriston Castle Collection. This collection is a subset of the larger library established by William Robert Reid, a prosperous Edinburgh businessman who acquired Lauriston Castle in 1902.[91] Reid had been assisted in his book collecting by a family friend, John A Fairley, author of several articles on the bibliography of chapbooks. During the course of his research Fairley had formed a collection of chapbooks containing around 500 volumes comprising over 5,500 items, which are now also part of the collection. The chapbooks are organised according to the town where they were printed, with the assortment mainly consisting of Scottish chapbooks, but English and Irish volumes also contribute to its diversity.

The NLS's Digital Scholarship Service is responsible for curating the digitised chapbook dataset which, along with other machine-readable data collections, is made available through the NLS's Data Foundry platform. While specific funding information is not provided, the NLS's Data Foundry operates as a permanent branch of the library, indicating that the resources were drawn from the library's overall budget rather than

---

[91] "Lauriston Castle Collection," accessed August 16, 2023, https://digital.nls.uk/catalogues/special-and-named-printed-collections/?id=598.

secured through separate funding endeavors. Further, the item-level data that is viewable after downloading the dataset from the Data Foundry platform explicitly states that, "NLS chapbook's digital form was curated as part of Library activities to make more Scottish collections available online, and chapbooks were selected for this task due to ease as all volumes are the same size."[92]

Although the collection process is not outlined in detail in the context of digitization, the NLS do offer information about how the chapbooks which were digitised were acquired. The Lauriston Castle Collection was bequeathed to the library in 1926 by Mr and Mrs Reid, following the latter's death that year. This bequest also included the Reid Fund, consisting of £70,000 (the income from the estate of Mr and Mrs Reid) which subsequently enabled the Library to acquire printed and manuscript items to add to the national collections.[93] To judge from the date stamp included in the Data Foundry's listing for the chapbook collection, the digitisation process presumably concluded in 2019. However, in the item-level data, it states that the chapbooks were captured as part of a project to digitise such materials beginning in 2015. The item-level data further reveals that each page of the chapbooks was captured using a Nikon D800E DSLR camera in the NLS's Causewayside studio by Picturae, a digitisation service provider. It is stated that transcriptions associated with each resulting image file were generated from optical character recognition (OCR) performed by the National Library of Scotland. There are many different algorithms and software packages for performing OCR, but

---

[92] Scotland, "Chapbooks Printed in Scotland".

[93] "Lauriston Castle Collection".

when we look more deeply at the item-level data, the images for each book were combined into a PDF format using a Luratech PDF application which has integrated OCR technology to make the PDFs searchable. In the page-level data, it is stated that pdfalto, a command line tool for parsing PDF files and producing structured XML representations of the PDF content in ALTO format, was used to breakdown the OCR output further.

Despite the straightforward reasoning provided surrounding the NLS's motivations for digitizing the chapbooks and the transparency of the digitisation process itself, what remains missing from the process of collection and curation is a question of *who*. The specific individuals responsible for collection and curation decisions are not documented. The identity of those who selected the items for digitisation is undisclosed. Additionally, information regarding the original ownership of the chapbooks before Fairley, the cataloging processes pre- and post-digitization, and the individuals involved in transforming the chapbooks into their current digital form within the NLS's Data Foundry remains missing from the available information.

The Archaeology of Reading

As touched upon earlier, AoR was created as a collaborative research endeavor to consolidate an exemplary portion of the marginalia produced by Gabriel Harvey and John Dee. The assembly of AoR began in November 2015 and was completed in January 2019, with its research and development being conducted with major funding from the Andrew W. Mellon Foundation. The books which make up AoR were selected and annotated by a collective of researchers at the Sheridan Libraries, Centre for Editing Lives and Letters, and Princeton University library, however, the lead researchers Earle

Havens, Anthony Grafton, and Lisa Jardine likely had the strongest curatorial role. The largest gap within the AoR collection is that they do not cover what technology was used to capture their texts, although they do indicate that the digitisation of the books was done primarily in situ by the repositories who held the physical copies themselves, or through a contract with UCL Digital Media. Further, they indicated that they required the images to have a resolution of 600 DPI which implies a DSLR camera like that used by the NLS. While they do not discuess what tools were use specifically, possibly because they themselves did not know due to the geographic expanse of their project, out of all digital collections used for this project, AoR provides the most extensive and transparent description of their collection process, largely in the form of a detailed article on how to "do" AoR yourself, that is, how to replicate their work using a researcher's own corpus, by closely explaining their own process.[94]

AoR is also the most prolific when it comes to discussing the curation of their collection, with dedicated essays on the libraries of Harvey and Dee. Although it is thought that Harvey's library once contained up to 4000 books, following his death it was dispersed with his books scattered in private, public, and academic libraries around the world.[95] So, the selection of which of his books to digitised for the AoR project was in large part a practical endeavour; the first books chosen were the nine in the possession of the Princeton University Library, one of their partnering institutions. In addition, these nine books, other titles were added to the Harvey AoR corpus often based on factors such

---

[94] "Archaeology of Reading", "How To Do AoR Yourself."

[95] "Archaeology of Reading", "A History of Gabriel Harvey's Library."

as availability (does the binding allow for the book to be digitised?) and the affordability of digitisation within a given repository. The Princeton books alone did not form thematic unity, as much as they reflected Harvey's intellectual interests in topics of warfare, (Roman) history, law, political economy (i.e., husbandry), and linguistics. However, the inclusion of five other titles alongside those at Princeton allowed for the expansion of the topics and the formation of "clusters" of books: books which thematically overlap, and which may have been read in conjunction with one another, as Harvey enjoyed doing.

While Dee's library was also dispersed posthumously, and in part, prior to his death due to financial troubles, he created a detailed catalogue of his books at numerous points in his life, which made his pursuits much more easily traceable.[96] Like when constructing Harvey's AoR corpus, factors such as the availability of books and the price charged by the various institutions for their digitisation were taken into account. However, with more choice being present due to the number of identifiable books annotated by Dee, further decision about what of his library should be digitise relied on intellectual interest. It was decided that primarily, the books selected from Dee should comprise of types and styles of reader interventions that are not represented in the Harvey corpus, as Dee's corpus contains several new interventions, including the use of additional symbols, genealogical trees, complex astrological charts, dense tables, and expansive drawings. Additionally, to further reflect a variety of reading and annotation strategies, the AoR Dee corpus also includes lightly annotated books such as Euclid's *Elementorum libri XV*, as well as different book formats, ranging from Cicero's *Opera* in

---

[96] "Archaeology of Reading", "A History of John Dee's Library."

folio to Gerhard Dorn's *Chymisticum artificium* in octavo. Lastly, in relation to Dee's library in its entirety, as in Harvey's corpus, the books included in Dee's corpus were selected to reflect the various intellectual interests which Dee pursued throughout his life, including mathematics, astrology/astronomy, medieval history, and New World discovery. The utmost goal when curating both corpora was an act of balance, reflecting the attempt to cover a representative selection of both readers' intellectual interests and the ways in which they interacted with their books.

Early Modern Annotated Books

The Early Modern Annotated Books collection hosted on Calisphere, a digital collections hosting platform for cultural heritage institutions based in California, was largely curated by the William Andrews Clark Memorial Library (the Clark) which is administered by the University of California (Los Angeles)'s Center for 17th & 18th Century Studies. This rare book and manuscript library specializes in the study of England and the Continent from the Tudor period through the long eighteenth century.[97] The digitisation of the Early Modern Annotated Books collection was initially a 2014 pilot project to digitize just ten annotated books from the Clark library, largely conducted by Philip Palmer who at the time was employed for a CLIR postdoctoral fellowship on the subject of "Manuscript Annotations in Early Modern Printed Books". A small grant from the Gladys Krieble Delmas Foundation allowed the ten books to be transcribed through the hiring of three graduate students onto the project, and the further digitisation

[97] "UCLA / William Andrews Clark Memorial Library," *Calisphere*, accessed August 21, 2023, https://calisphere.org/institution/62/collections/.

of annotated books within the Clark's collections was made possible through funding from the National Endowment for the Humanities, which awarded the library a Humanities Collections and Reference Resources Grant in 2017.[98]

The collection process of the William Andrews Clark Memorial Library lacks clarity; the metadata is largely bibliographic, and some books include a section on provenance, but it is the provenance of the physical item rather than the digital. There is no specific information on when the Early Modern Annotated Books collection was assembled in both physical and digital form. Palmer states that the process of digitisation began in 2014, however when referencing the funding statement given in the collection's official description, it is implied that the books were captured during the time which the National Endowment for the Humanities grant was held, between 2017 until the project's end in October 2018. There is no clear information provided on the tools used by the Clark for digitization, although looking at the EXIF data extracted by my annotation application, there are tags such as GPS common in TIFF images that may point to a camera having been used over technology like a scanner. In contrast to the NLS and AoR, there is also very little information about the decision-making process made in the curation of Early Modern Annotated Books collection. Palmer selected the first ten books to be digitised for the collection based on how they were "representative of the characteristic idiosyncrasy that historical readers brought to their material readings of

---

[98] Philip Palmer, "Annotated Books at UCLA: Wider Applications of the AoR Schema Archaeology of Reading," *Archaeology of Reading* , September 2018, https://archaeologyofreading.org/annotated-books-at-ucla-wider-applications-of-the-aor-schema/.

books", however it is unclear exactly which ten books these are.[99] Based on those which he discussed in his 2018 blog post on the project, this ten may have included a copy of Sir Thomas Browne's *Pseudodoxia epidemica*, a copy of the 1603 English translation of Montaigne's *Essayes*, Richard Allestree's *The Art of Contentment* (1675), Aleazar Albin's *The Natural History of English Song-Birds* (1779), Sir Richard Blackmore's *Prince Arthur*, and Voltaire's *Dictionnaire Philosophique*.

Evidently, the Clark's Early Modern Annotated Books collection is the collection which leaves the most unknowns, and this seems to be at least in part due to the focus on the digitised books as being surrogates for the physical object, rather than a "new edition".[100] The metadata associated with each book appears to be about the physical book or where the physical book is within the holding institution, and this notion is affirmed by the Calisphere "statement on digital primary resources".[101] In this statement, these digitised entries are referred to as primary sources themselves and in the section discussing their metadata, they are discussing metadata created from cataloguing the physical item rather than from the digital. It seems that this may not have always been the case, given the pilot project which started the Early Modern Annotated Books collection attempted to incorporate elements of the AoR project's XML schema for the original ten books, however this effort seemed to be abandoned once the project moved past the pilot, and even the XML files and transcribed annotations that did exist appear to be no more,

---

[99] Palmer, "Annotated Books at UCLA".

[100] Cordell, ""Q i-Jtb the Raven"".

[101] "About the Collections in Calisphere," *Calisphere*, accessed August 21, 2023, https://calisphere.org/overview/.

with the link provided in the already obscured blog post describing this process being broken.[102] The neglect of the pilot project that originated this collection is also an issue in itself. There is no evident reference to Palmer's work on developing the project on any of the institutional platforms that define the archive, yet all in-depth information on this collection is derived from a blog post by Palmer written in 2018, which I found by happenstance since this post was a guest publication on the AoR website's blog section. The link to any discussion on how this digital archive came to be is severed to those looking at the collection as it is presented on Calisphere. Further, all additional content crafted by Palmer being no longer available seemingly indicates that the Clark made no effort to preserve these original components of their present collection.

Dataset Provenance

Although all digital collections used for this case study fell into the public domain and allowed their content to be used for research purposes, neither AoR nor the Clark's Early Modern Annotated Books collection provided their data in an easily downloadable format for researchers who wish to work with their collections computationally. AoR does attempt to offer the relevant data via data releases throughout the project's development, however these data releases contain only the project metadata and not the images needed for the purposes of annotation. In consideration of these limitations, I chose to take the common approach of webscraping to collect the images and associated metadata needed to build my training dataset. Webscraping is the automated process of extracting information and data from websites; it involves using digital tools to gather

---

[102] Palmer, "Annotated Books at UCLA".

and parse through web pages, collecting data based on parameters set by the person using the scraper.

Using the Python programming language to write the scripts to perform webscraping, the general structure I followed for webscraping was first, gathering the links to each digitised text's entry, then iterating over each page of the text to extract the image from the webpage and save it to my device. While this approach worked broadly, each archive had its own intricacies that required customization of the webscraping code. AoR uses Mirador, an all-encompassing viewer for exploring and interacting with digital objects and collections of cultural heritage materials, to display each book in their collection, the images are difficult to extract from the webpage's HTML. To deal with Mirador, a simple webscraper would not work. Instead, I had to write code that would mimic a human being paging through the results and right-clicking 'save as', over and over again. This was accomplished with the Python library Selenium, which allows for the automation of web browser interactions– essentially, mimicking the actions a person might to perform a task should it be done manually. In this case, for each book I simulated the process of hitting right-click and "Save As" on each page image that was indicated as containing marginalia based on the transcription metadata AoR provided, entered what I desired the file name for the downloaded image to be, which in this context consisted of the page number followed by the book label (ex. 15-MattheusBeroaldus-Chronicum(Geneva-1575).jpg), then downloaded the image and hit the arrow button which would lead to the next page, where this process would then repeat until all pages had been downloaded. Since the Clark's Early Modern Annotated Books collection is much larger than AoR, I firstly added the constraint that only books with a

page count less than 450 should be downloaded to avoid overloading the storage on my device. Then, I was simply able to download each page image through extracting the link associated with the "Download Image" button present on each page entry.

As discussed, because the process of annotating images inherently removes them from their intended context which can contribute to their misuse, the metadata associated with each page was scraped alongside the images and appended as EXIF data so that each page would carry its archival context with it through the process of annotation. The metadata is transformed into a Python dictionary upon extraction from the webpage's HTML, then this dictionary is transformed into JSON format. This structured data was attached to the image via the EXIF data's UserComment field, and then the application extracts JSON data present in the UserComment field.[103] This makes it so the metadata is viewable in-app, but also within the app's save file, which is simply a JSON file.

Only the materials available through the NLS's Data Foundry were made available from the beginning with the intention that they be used computationally. However, returning to the discussion of cultural heritage institutions appropriately formatting data for the intended audience, the Data Foundry distributes the metadata for their datasets as METS files at the item-level and ALTO XML files at page-level. Although the information contained within these files is valuable, these formats are very dense and difficult to parse for readers, human and computer alike, attempting to gather information from them. I attempted to use Python to parse the metadata NLS provides in

---

[103] This is done using Python's json library, as well as Python's exif library for .jpg images and the PIL library for .png images.

these METS and ALTO files into the JSON format needed in order to be manipulated and used by the application, but this proved more difficult and required more time and energy than was perhaps warranted; just because something is computationally possible does not mean that it is necessarily easy to achieve. As with any other method, decisions must be made. Instead, as a compromise, I chose to rename each page to include the page number and the item-level reference number alongside the page-level reference number that was already present, so that the metadata files for each image could easily be found by searching the relevant reference number within in my file system.

Results

Brief Overview of Training the Machine Learning Model

Within the time frame I allotted to the task of creating an annotated dataset from the gathered collections, 353 early modern book page images were annotated in RocketAnnotator, then 20 examples of pages which lacked marginalia were added in. This dataset was then randomly split into an 85/15 training/test ratio, meaning that around 317 (~85%) images are allocated to training the machine learning model, 39 (~10%) images are used to validate the model while it trains, and 17 (~5%) images are reserved to test the results of the trained model. For effective object detection training, it is recommended to have 1500 images per new class, so to bolster my smaller dataset I added in augmented versions of the 317 training images to the training data. This is a common technique used to introduce variation in small dataset that can enhance the model's ability to generalize. Each image was replicated with intentionally added noise, contrast, brightness, or saturation adjustments, resulting in a total of 1585 training

images. However, even with these augmentations, the dataset remains small at 795 MB in the context of machine learning.

For the purpose of object detection, I chose to use the YOLOv7 model.[104] In general, the "You Only Look Once" family of models function by processing the entire image just once to detect objects instead of iteratively analyzing an image multiple times at different scales to identify objects as other object detection models have traditionally done, which makes YOLO models perform the task of object detection faster and more efficiently compared to other models. My dataset is small, so rather than training a model from scratch which would likely yield poor results due to this limitation, I chose to use transfer learning with one of YOLOv7's pretrained models. Transfer learning is a machine learning technique where a model that has been trained on one task is repurposed or "finetuned" for a different but related task. Instead of starting from scratch, the knowledge gained from solving one problem is transferred to help solve a different problem; this not only saves time and resources since the model is not being built from the ground up, it also builds upon what the model has already learned about recognizing characteristics such as shapes and patterns, making it more efficient and effective at expanding this palette.

With this method, the trained model yielded 77.6% precision, 79.5% recall, and 77.1% mAP@.5. Precision is the ratio of true positive detections to all positive

---

[104] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors," *arXiv Preprint arXiv:2207.02696*, 2022, https://doi.org/10.48550/arXiv.2207.02696.

detections, recall assesses the model's broad object detection accuracy, and mAP compares annotator and model drawn annotation boxes, determining successful detections. YOLOv7 specifically uses mAP@.5, where the .5 refers to the Intersection-over-Union (IoU) threshold for correct detection overlap; if the model's annotation box overlaps at least 50% with the annotator's, this is considered a correct detection. Thus, when my model predicts marginalia, it is generally correct, and the model is able to find and capture a substantial portion of the actual marginalia. Considering the small size of the training dataset and the goals of this project as a whole, these results are acceptable enough that the model should provide reasonably accurate detections of marginalia when applied to the NLS chapbook dataset.[105]

The Model's Detections

Of the 47329 pages present in the NLS chapbook dataset, a total of 10560 pages were detected as containing marginalia. Of these pages, 4239 actually contained marginalia (true positives), whereas the remaining 6321 contained aspects of pages which the model mistook as being marginalia (false positives). This significant number of false positives is in large part due to my decision to set the confidence score to 0.05 or 5% when using the model on the chapbook pages. The confidence score in the context of detection means that when the model is searching the pages for marginalia, if the model is at least 5% certain that something it found is marginalia, it will draw a bounding box

---

[105] For a more detailed discussion on the annotation and training process, see Appendix C.

around it. I set the confidence score lower to ensure no marginalia was missed given the lower precision score of the model.

Table 1. True Positives Categorized. **Total True Positives:** 4239.

Category	Count	Definition
Archival Traces	3736	Notes clearly denoting the archival life of the chapbook prior to digitization
Numbers	131	Numeric marginalia with no clear link to archival purposes
Uncertain	111	Primarily marginalia which is illegible due to ink, fading, font etc
Graphical Reading Systems	103	Symbols which denote active reading
Marks of Ownership	60	Names or initials
Corrections	51	Interaction with the text that primarily comprises of editing the contents
Text Interaction	31	Commentary on the text or about it
Mathematics	7	Math equations
Pen Trials	7	Scribbles or letters which serve to test the writing tool
Inserted Notes	2	Notes inserted into chapbooks

The most surprising outcome from these detections is how few of the pages actually contained marginalia penned by early modern readers. A vast majority of the pages with detected marginalia show instead the archival lives of the chapbook. Some of the earliest archival marks are potentially early modern in a generous sense; there are a number of chapbooks with “No. __” penned with ink in a cursive hand on the title page, and typically also found on these pages is a stamp which reads “Edin. S.S. Library”. This is the mark of Edinburgh Select Subscription Library, a private subscription-based library where the subscription was not only to borrow books but to hold a share in the library, creating a library which was owned by its shareholders. The Edinburgh Select Subscription Library was founded in 1800 by a group of ten young men to rival the

earlier established Edinburgh Subscription Library, who they felt had too high of subscription fees.[106] Aside from these marks, the remaining signs of cataloging are more contemporary. There is a handful of descriptive notes within the chapbooks which appear to be from the 20th century as are dated with the year 1910. This is likely the work of Fairley himself considering this would only be around 16 years prior to the NLS's acquisition of the Lauriston Castle collections. The most recent of the archival notes detected are numbers penciled in the upper corners of some of the images, on a paper placed behind the chapbook page rather than on the page itself. These numbers correspond to the final digits present in the NLS shelf mark of each chapbook, so presumably these are present in the images as reference for the pages being scanned.

---

[106] K. A. Manley, "Scottish Circulating and Subscription Libraries as Community Libraries," *Library History* 19 (July 2013): 191, https://doi.org/10.1179/lib.2003.19.3.185.

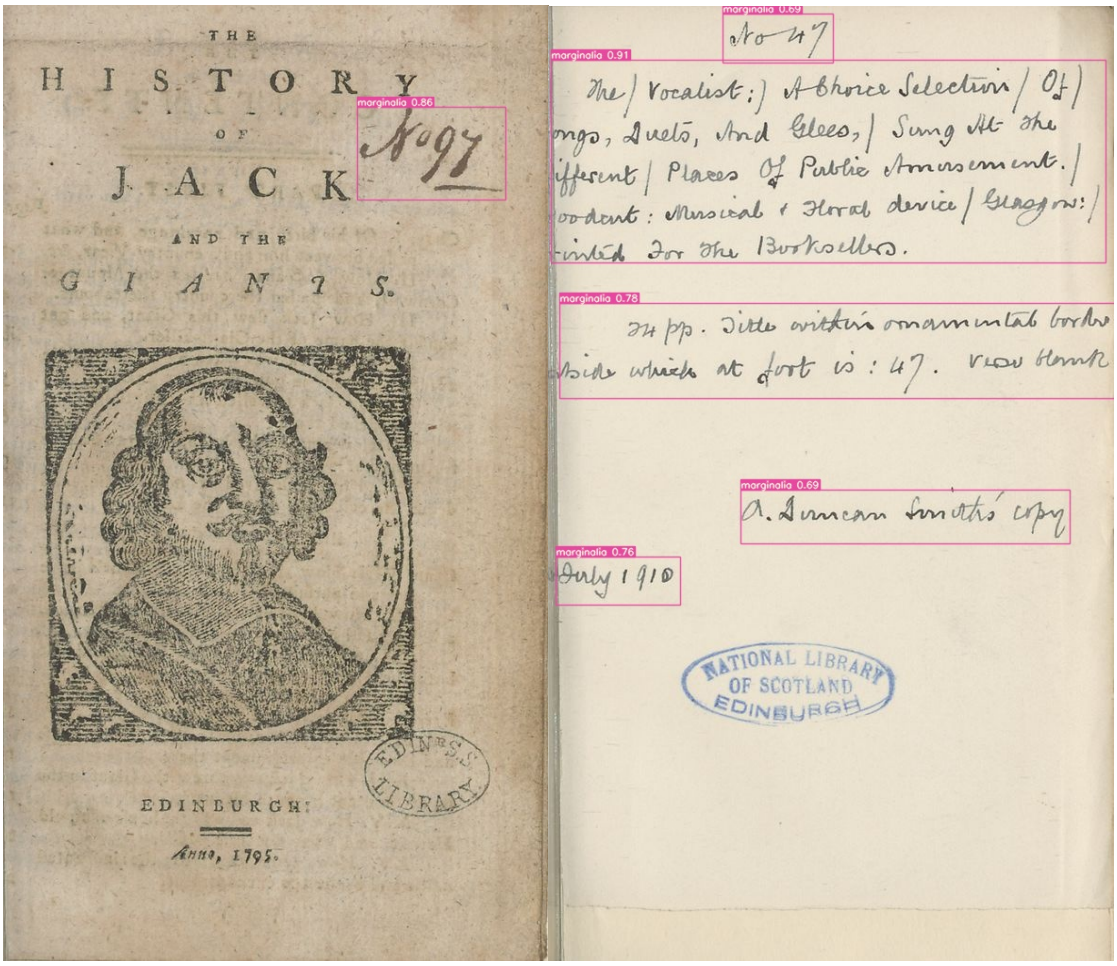


Figure 22. **Left:** Edinburgh Select Subscription Library marks. *The History of Jack and the Giants*, 1795, National Library of Scotland, http://digital.nls.uk/104184561. **Right:** Contemporary notes possibly belonging to John A Fairley. Printed for the Booksellers, *The Vocalist's Song Book*, ca. 1840, National Library of Scotland, Associated metadata file from Data Foundry: 104185187-mets.xml.

Notable outside of archival reference are the numerous examples of marks of ownership. The model uncovered even more of John Watson's exuberant signatures alongside new frequent signers, such as William Smitton who neatly placed his mark at the top of each of his ten chapbooks' title pages. Slightly less common but still notably present is another reader named Peter Smitton, whose signature appears in a similar placement to William's within older chapbooks seven times, revealing perhaps a generation of chapbook consumers. There is also occasionally evidence of female

readership, with a reader named Margaret Cameron labelling her three books in a manner similar to the Smittons (See Figure 2).[107]

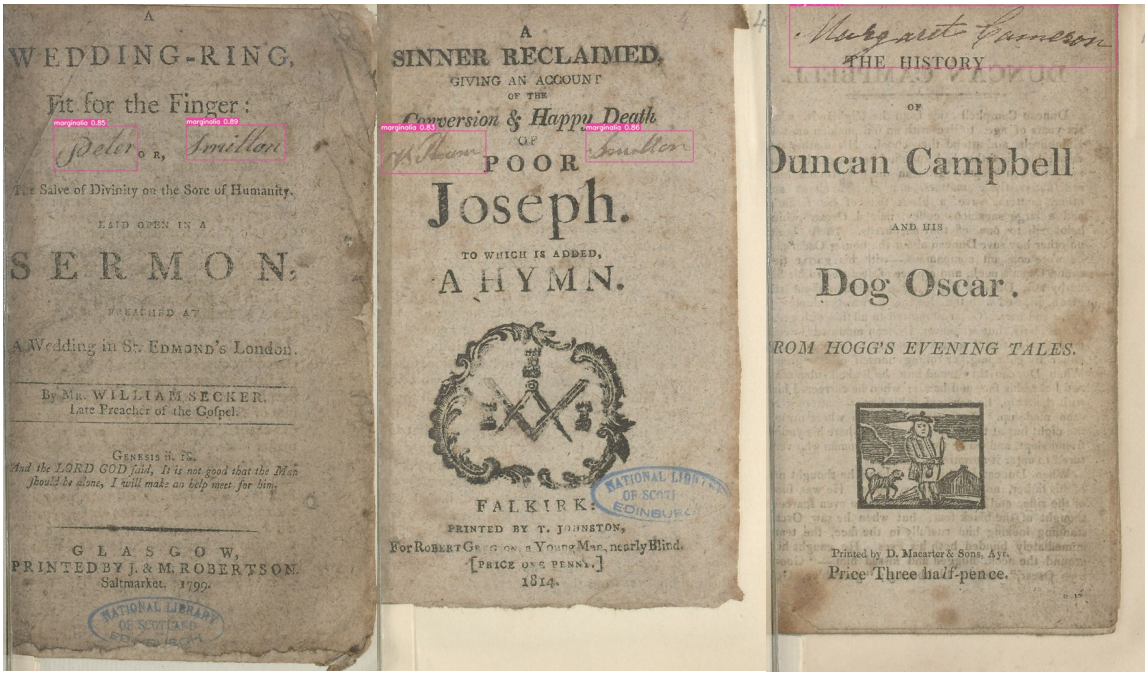


Figure 33. **From left to right:** Marks of ownership by Peter Smitton, William Smitton, and Margaret Cameron.

Beyond marks of ownership, there is evidence of the chapbooks being used for more pragmatic tasks. One of the first detections the model made and only one of this type was of what appears to be tally marks, perhaps someone keeping count of a task or a game or other common repetitive task, such as transactions. I propose game primarily,

---

[107] Reference for the pages in Figure 2 in order: Printed by J. & M. Robertson, *A Wedding-Ring, Fit for the Finger*, 1799, National Library of Scotland, http://digital.nls.uk/104185291; Printed by T. Johnston for Robert Gregon a young man nearly blind, *A Sinner Reclaimed*, 1814, National Library of Scotland, http://digital.nls.uk/104185116; Printed by D. Macarter & Sons, Ayr, *The History of Duncan Campbell, and his dog Oscar*, ca. 1817, National Library of Scotland, http://digital.nls.uk/104184176.

because the tally marks seem to be accompanied by some unusual scrawlings beneath, possibly a rudimentary attempt at spelling by a child. There were also a handful of more sophisticated examples of mathematics detected within the chapbook pages, with 7 sequences of multiplication and addition detected, being performed for an unspecified task. There are examples of similar calculations being performed in the Early Modern Annotated Books collections in the almanac-turned-account book of an 18th century wigmaker, so perhaps these chapbook calculations were also related to business, personal or professional.[108]

[108] Edmund Weaver, *Wigmaker's Account Book.* (London: A. Parker for the Company of Stationers, 1737), https://calisphere.org/item/ark:/21198/n14s4d/.

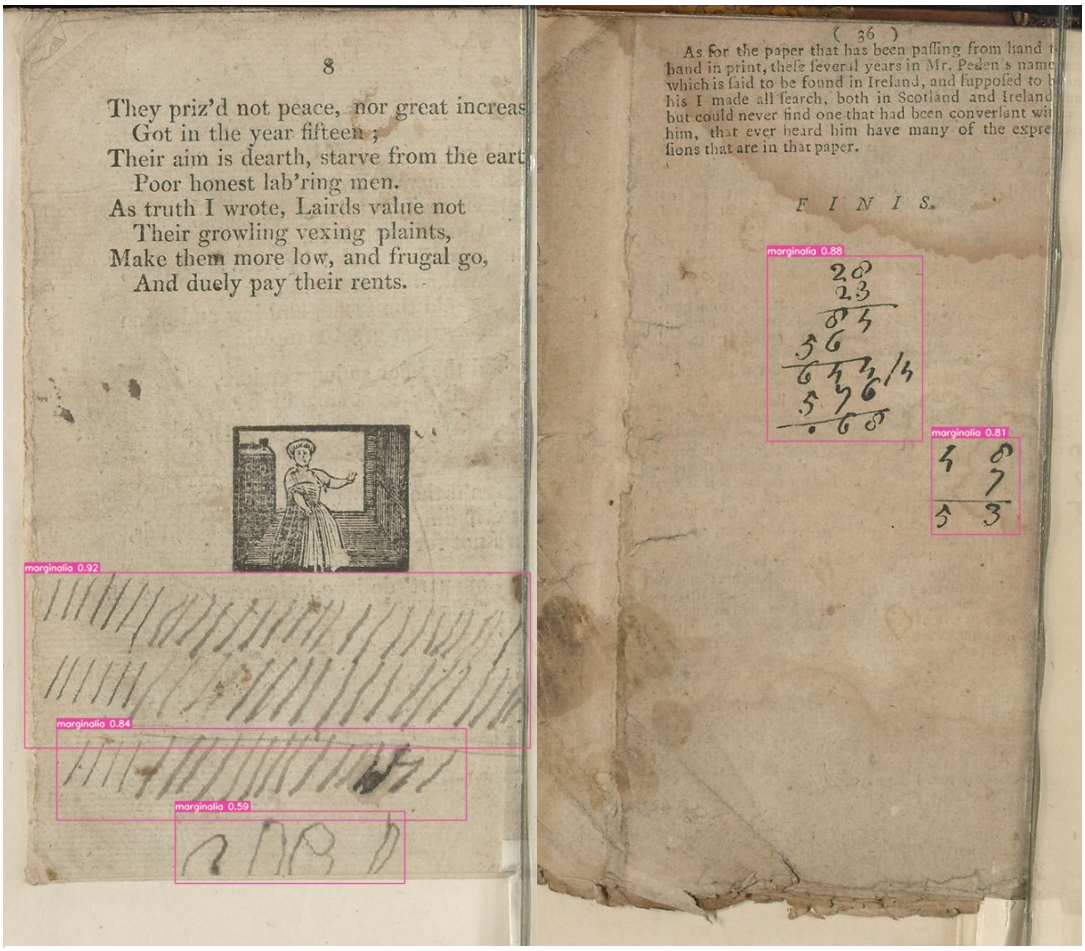


Figure 44. **Left:** Mysterious tally marks. Printed by D. Macarter & Co, *The Historical Ballad of May Culzean*, ca. 1817, National Library of Scotland, http://digital.nls.uk/104184172. **Right:** Example of math equation. *Some Remarkable Passages of the Life and Death of Master Alexander Peden*, 1760, National Library of Scotland, http://digital.nls.uk/104185268.

There were also marginalia detected that demonstrated engagement with the chapbook's text. There are a number of examples of light annotation which might fall under what Grindley classified as "Narrative Reading Aids", with notes clarifying the meaning of words or phrases the reader seemingly did not initially understand; for example, one reader upon coming across the term "Whip whire" wrote below it, "Bird". There are also examples of literary response, primarily in the form of correcting and expanding the chapbook's text. Moreover, graphical responses using systemised forms of graphic shorthand or added punctuation are plentiful, with crosses (+) and x marks being

left in places which the reader deemed notable or significant to their understanding of or connection to the text.

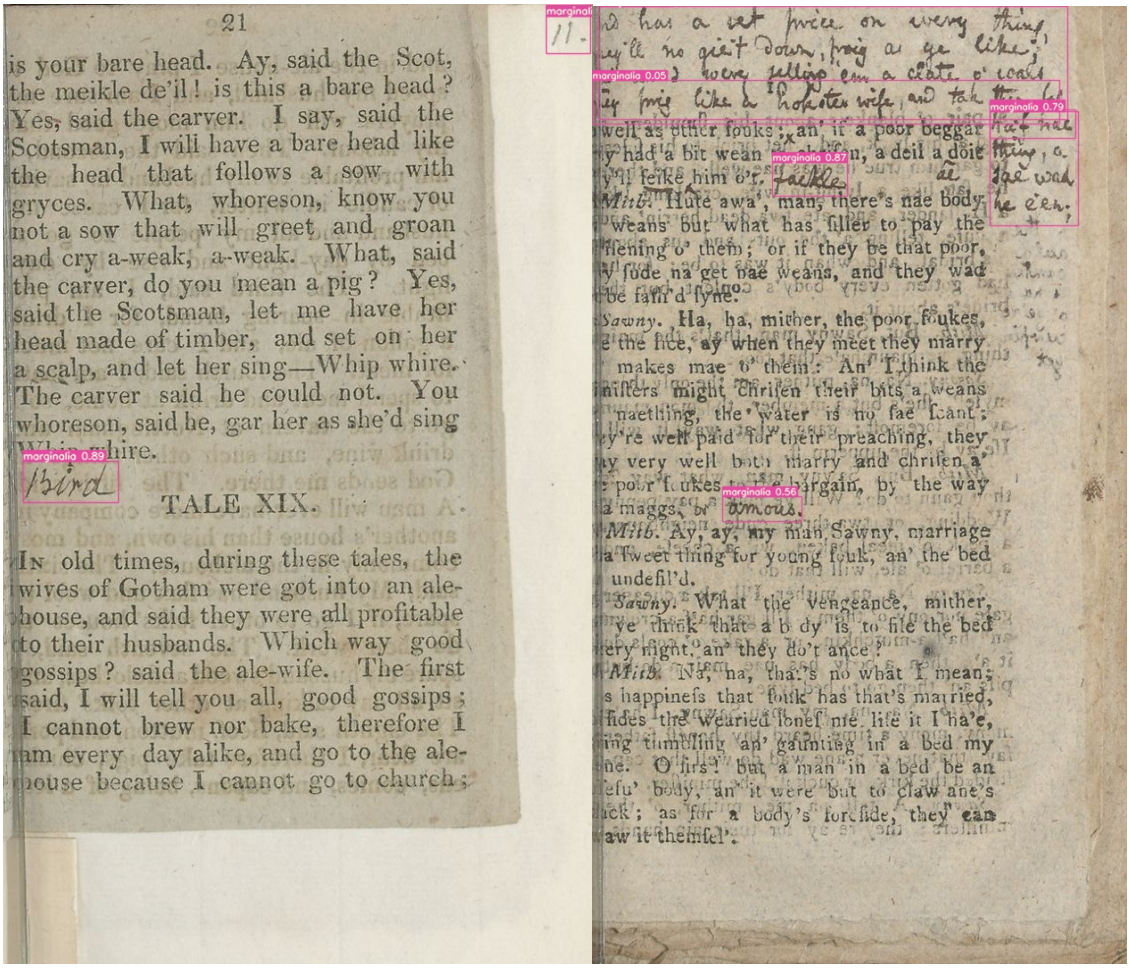


Figure 55. **Left:** A note clarifying the meaning of 'Whip whire'. Printed for the booksellers in town and country, *The Merry Tales of the Wise Men of Gotham*, ca. 1815, National Library of Scotland, http://digital.nls.uk/104184792. **Right:** Interactions with the text. *The Coalman's courtship to a Creelwife's Daughter*, ca. 1799, National Library of Scotland, Associated metadata file from Data Foundry: 104186983-mets.xml.

False Positives

Table 2. False Positives Categorized. **Total False Positives:** 6321.

Category	Count	Definition
Print flaws	2815	Poor printing quality resulting in random ink marks on page which model mistook for marginalia
Illustrations	1169	As implied
Ink Bleeding	767	Ink bleeding through pages forming what appears to be new marginalia
Mistaken reading systems	686	Symbols within printed text near places which annotators have used symbols in training data
Font	551	Rough or italicized section of printed text which results in model mistaking irregularities for handwriting
Paper quality	266	Poor paper quality with visible fibers and significant creasing causing shadows on page which model mistake for marginalia
Ink spills	33	Ink splatters on page
Printed tables	18	In training data, tables are largely hand drawn, so model mistook printed tables for hand drawn ones
Hand traces	16	Thumb prints

Within this collection of pages that the model detected as containing marginalia, there were many false positives which is expected given the middling evaluation metrics outputted for this model. However, the patterns found within these false positive detections offer clear insight into where the training dataset could be strengthened. The model detected ink smudges from hands, ink bleeds through the pages of the paltry paper, ink spills, and printing errors as marginalia– all elements which demonstrate the rough process of production and the human touch which chapbooks underwent in their early lives. Although there were certain pages that had been weathered with age in the training dataset, in general, the pages in the training dataset were much cleaner than the chapbook pages, due to what appears to be higher quality paper and more careful printing methods.

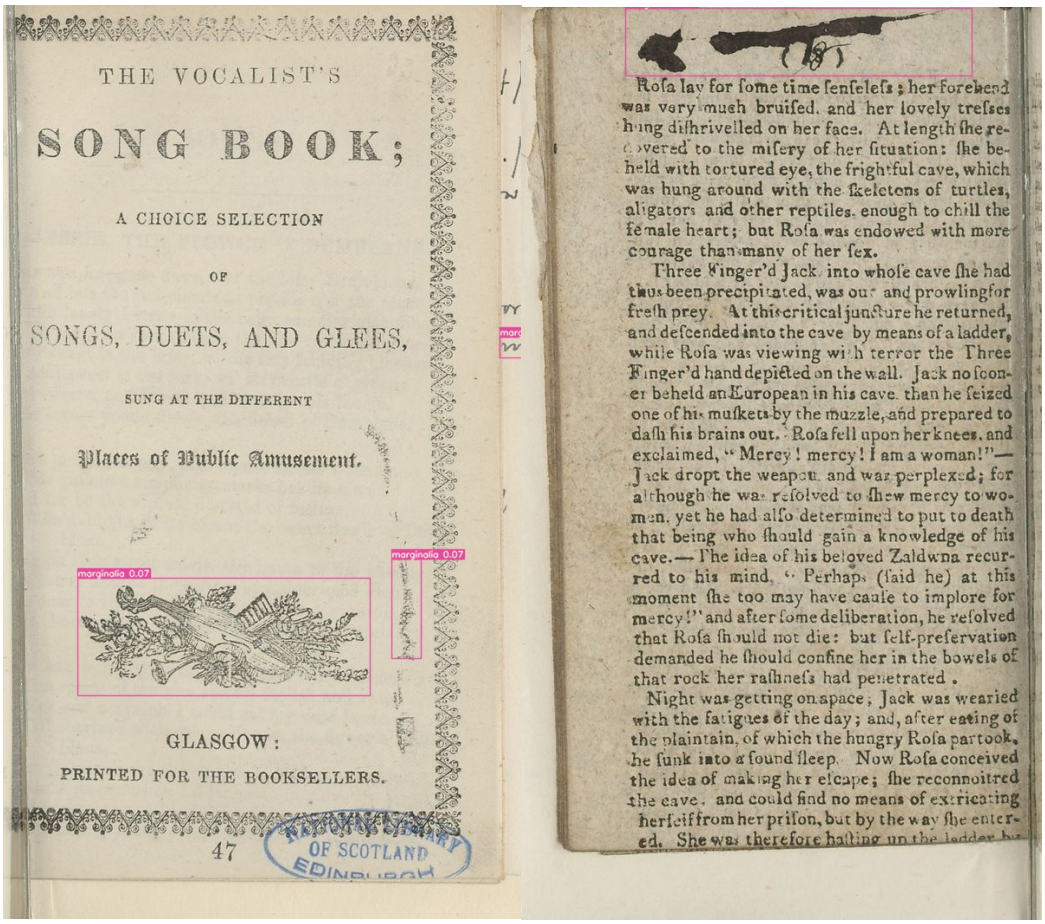


Figure 66. **Left:** Flaws from printing press. Printed for the Booksellers, *The Vocalist's Song Book*, ca. 1840, National Library of Scotland, Associated metadata file from Data Foundry: 104185187-mets.xml. **Right:** Example of an ink spill. Printed and sold by T. Johnston, *The History and Adventures of Three Finger'd Jack*, 1822, National Library of Scotland, http://digital.nls.uk/104185025.

Roughly printed punctuation marks present near the page margins were often detected as marginalia, likely due to the way their placement and appearance is similar to the systemised forms of graphic shorthand employed by readers Dee and Harvey, who were both heavily featured in the training data (see Figure 6). Likewise, sections of printed text which were italicized, unclear due to imprecise printing, obscured by the text on the previous page seeping through, or any combination of these factors would occasionally be misdetected as being handwritten. These flaws showcase a level of

irregularity that the machine understands as being characteristic of handwritten rather than printed text.

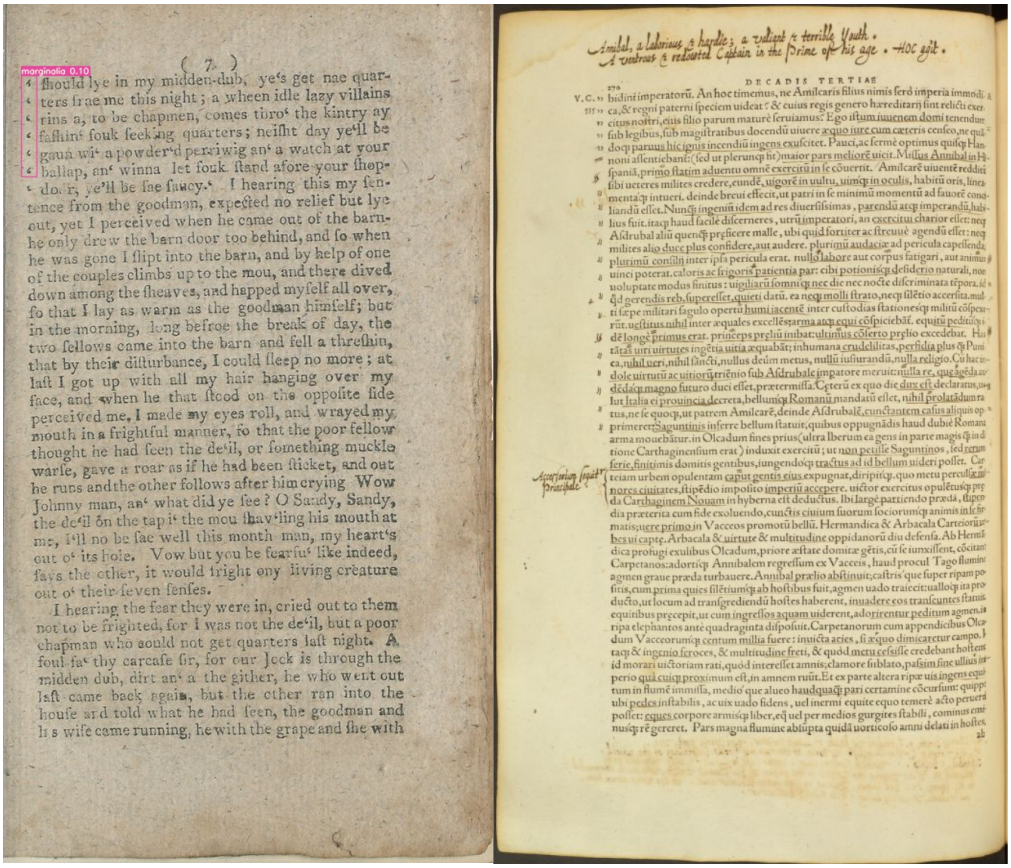


Figure 77. **Left:** Mistaken punctuation. Printed and sold by John Morren East Campbell's Close, Cowgate, *The History of John Cheap the Comical Chapman*, ca. 1800, National Library of Scotland, http://digital.nls.uk/104184596. **Compare to right:** Harvey reading system. *Titus Livius, Romanae historiae principis*, 1555, Princeton University Library, https://archaeologyofreading.org/viewer/#aor/PrincetonPA6452/binding.frontcover/image.

For related reasons, the model also extensively detected the woodcut illustrations and page decorations within the chapbooks as marginalia. This is a particularly notable flaw in my dataset, as upon revision of the images annotated, there are very few examples that included illustrations on the page alongside marginalia meaning that the model never learned to fully differentiate between these types of "free form" shapes. In this same vein, within the training data there was a small number of hand-drawn tables, thus the model was able to gain an understanding of the tables' shape but not necessarily the features

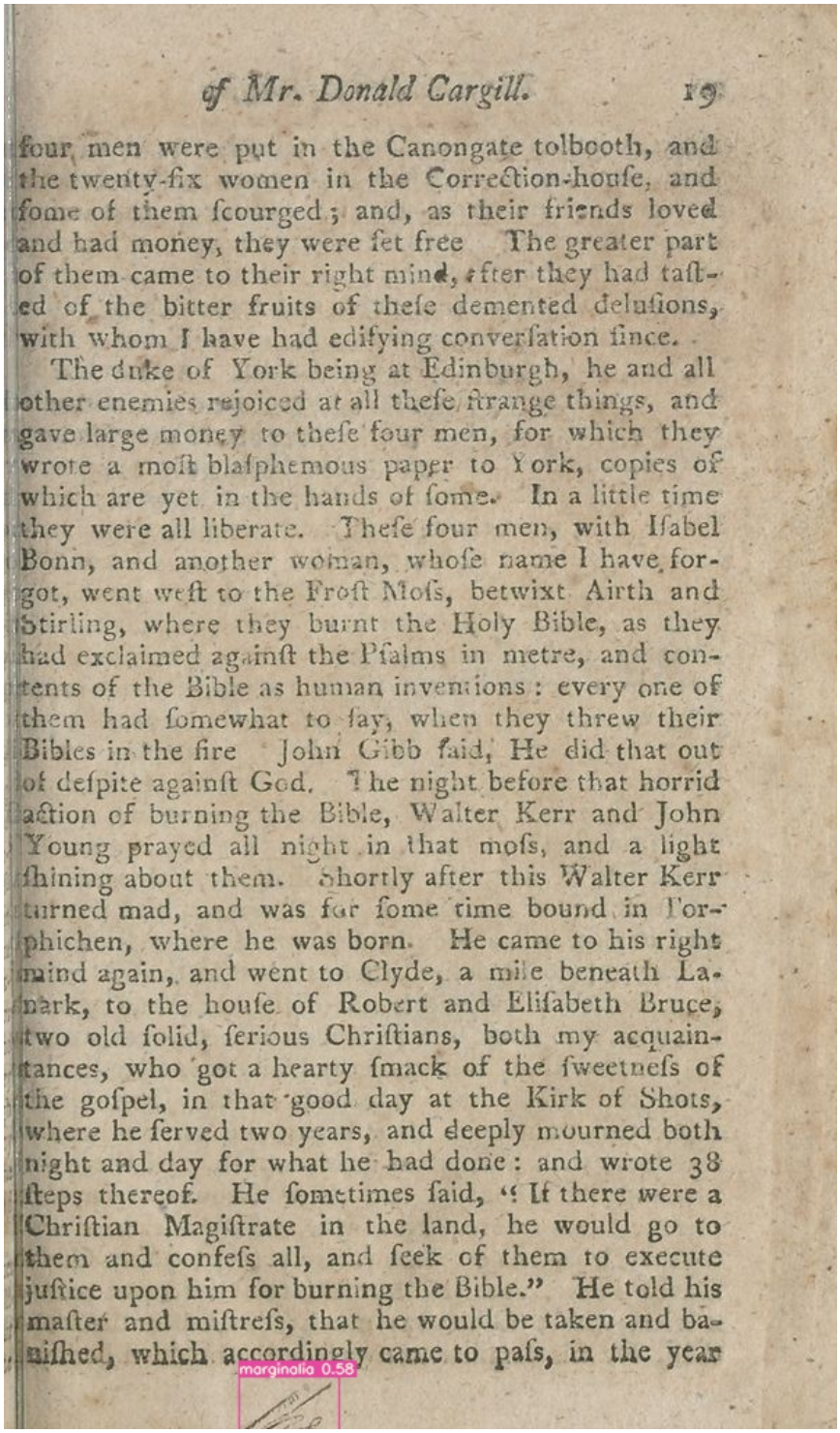
which distinguished these tables as being hand-drawn, resulting in the printed tables present in the chapbooks being detected as marginalia.

Human vs the Machine

When comparing the performance of this model to that of a human when it comes to locating marginalia, the first thought may be that a human would perform better, as we would not create the false positives that the model did. It is true that the human brain is much better at identifying objects it may have seen only a handful of times– but, to manually perform the task that the model did, we would have had to look at each page of all the chapbooks present in the NLS collection one by one. When discussing his manual process of finding and studying marginalia for *Used Books*, Sherman indicated that his work had taken over a decade to complete.[109] In contrast, the model took approximately 1.5 hours to iterate over and detect marginalia across the entire collection of chapbooks, then it took me around 8 hours to go through the outputted pages and sort them into categories. The goal for this case study was not necessarily to create a model that is better than humans (although it certainly could be improved to be closer to our ability), but to create a tool that makes the process of finding and by extension studying marginalia much faster/more efficient. Although, in saying this, the model did perform well when it came to finding marginalia that I feel I likely would have missed; for example, in a

[109] Sherman, *Used Books*, xvii.

chapbook otherwise devoid of marginalia there appears to be one page that has either a possible mark of ownership or pen trial peeking out from the edge of the page.



of Mr. Donald Cargill. 19

four men were put in the Canongate tolbooth, and the twenty-ſix women in the Correction-houſe, and ſome of them ſcourged; and, as their friends loved and had money, they were ſet free. The greater part of them came to their right mind, after they had taſted of the bitter fruits of theſe demented deluſions, with whom I have had edifying converſation ſince.

The duke of York being at Edinburgh, he and all other enemies rejoiced at all theſe ſtrange things, and gave large money to theſe four men, for which they wrote a moſt blaſphemous paper to York, copies of which are yet in the hands of ſome. In a little time they were all liberate. Theſe four men, with Iſabel Bonn, and another woman, whoſe name I have forgot, went weſt to the Froſt Moſs, betwixt Airth and Stirling, where they burnt the Holy Bible, as they had exclaimed againſt the Pſalms in metre, and contents of the Bible as human inventions: every one of them had ſomewhat to ſay, when they threw their Bibles in the fire. John Gibb ſaid, He did that out of deſpite againſt God. The night before that horrid action of burning the Bible, Walter Kerr and John Young prayed all night in that moſs, and a light ſhining about them. Shortly after this Walter Kerr turned mad, and was for ſome time bound in Torphichen, where he was born. He came to his right mind again, and went to Clyde, a mile beneath Lanark, to the houſe of Robert and Eliſabeth Bruce, two old ſolid, ſerious Chriſtians, both my acquaintances, who got a hearty ſmack of the ſweetneſs of the goſpel, in that good day at the Kirk of Shots, where he ſerved two years, and deeply mourned both night and day for what he had done: and wrote 38 ſteps thereof. He ſometimes ſaid, "If there were a Chriſtian Magiſtrate in the land, he would go to them and confeſs all, and ſeek of them to execute juſtice upon him for burning the Bible." He told his maſter and miſtreſs, that he would be taken and baniſhed, which accordingly came to paſs, in the year

marginalia 0.58

Figure 88. Marginlia peeking over the image edge. Printed by J. Neilson, *The Life and Prophecies of that Faithful Minister of God's Word, Mr Donald Cargill*, 1812, National Library of Scotland, http://digital.nls.uk/104186348.

It seems that this page was poorly cropped following digitization, so at first glance these fine pen marks look as if they could be another artifact of printing or fiber in the paper and be passed over if the researcher is not intensely focussed on the task at hand. One advantage that machines *do* have over humans is that we get tired by repetitive tasks and in response may become inattentive or start rushing to finish, whereas machines do not experience this.

It should be noted that this study of false positives is where the application showed its use beyond just the image annotation functionality. I was easily able to reflect upon the training data and identify the source of these false positives through simply opening the annotation project in the application. I was able to revise the quality of my annotations and what my training data lacked in comparison with how the model performed. Having an easy way to view and search my clearly recorded training data gave me the ability to see the potential my model has for improvement.

Discussion

One categorical observation that can be made about these marks broadly is that when comparing the marks of ownership which appear to belong to children as indicated by their larger and less constrained hand writing, to those of adults, it seems that children often like to assert firm claim upon the chapbook by appending statements such as “his book” or “is my name” to their signature, occasionally alongside a misspelled date. In his research on children’s marginalia, scholar Seth Lerer identifies these as “stories of possession”— notes which clearly define who the text belonged to, protecting an object

perceived as important by the owner.[110] Evidence of what appears to be children sharing did however, also show up among the detections, with one page of a chapbook bearing three different names all in different hands and pens, the chapbook seemingly being used as hand writing practice for a group of friends or siblings.

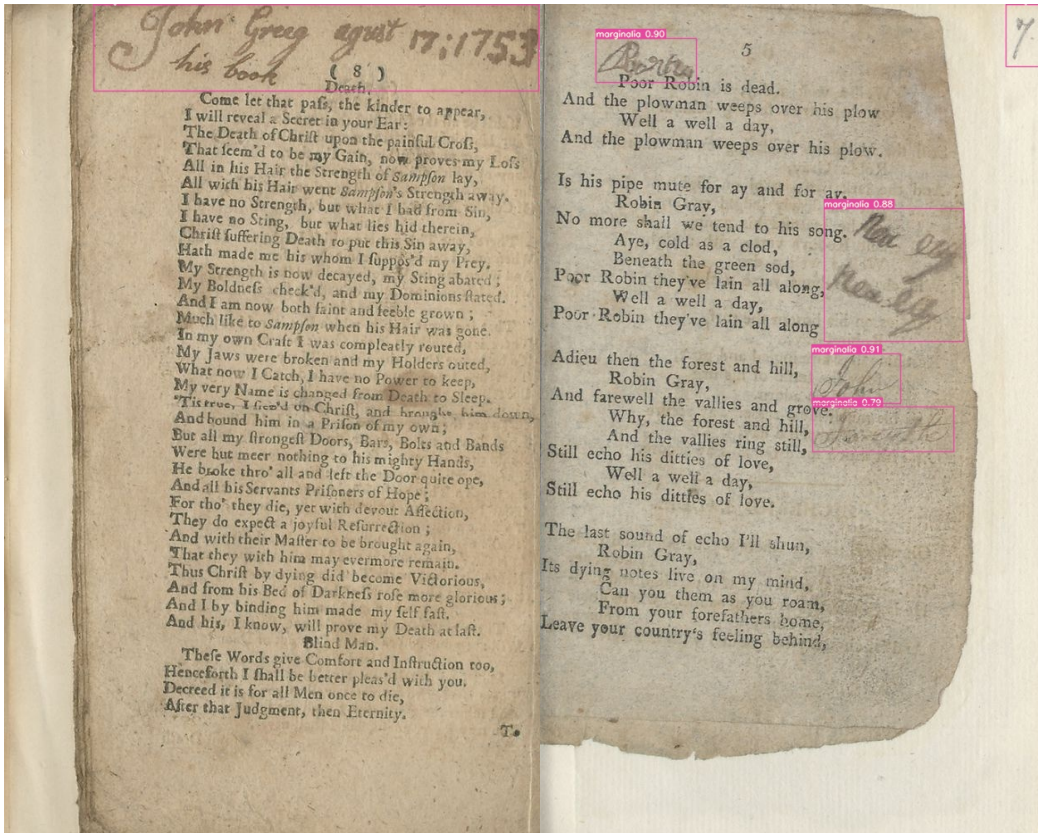


Figure 99. **Left:** A book claim. Printed and sold in Swan-Close, *A Dialogue between a Blind Man and Death*, ca. 1740, National Library of Scotland, http://digital.nls.uk/104184326. **Right:** A collection of signatures. Printed for the booksellers, *Birks of Aberfeldy*, 1817, National Library of Scotland, http://digital.nls.uk/104184636.

I began this project believing that chapbooks would be the perfect medium for marginalia of the "common" class, given how prolific yet affordable these little books were. Yet the results from the model made me question, why is there so few pages with

[110] Seth Lerer, "Devotion and Defacement: Reading Children's Marginalia," *Representations* 118, no. 1 (2012): 135, https://doi.org/10.1525/rep.2012.118.1.126.

marginalia present? Observing the chapbook marginalia in its entirety, outside of the content written on the page, an interesting history of the object emerges through a way in which the physical properties of the chapbook's materiality become apparent despite digital format via the writing mediums used by the annotators. In his entry within *Early Modern English Marginalia*, scholar Joshua Calhoun introduces the topic of gelatin sizing, the viscous gelatin solution in which paper was dipped during the early modern period to render it suitable for writing with the water-based ink used for manuscript.[111] Conversely, this discussion also brings up the topic of poorly sized paper and "sinking", a contemporary term used to describe paper that could not hold its ink; using porous paper would cause ink to spread, absorb, or run on being applied to it.[112] Chapbooks, evidently, were printed on paper that was at most poorly sized, and the marginalia clearly illustrate this. Many marginalia such as that of John Watson's look as if they were written with water colours due to the way the ink spread on the paper, and even the clearest marginalia still suffered from some bleeding along the edges and the occasional blob of ink, obscuring what is written. There are multiple instances of detected marginalia sinking so severely that their meaning is blotted out completely. While chapbooks may seem to be an ideal medium for quick notes and scrap paper given their low cost and proliferation, perhaps such a small percentage of the pages actually contain marginalia in practice since the construct of the paper did not lend itself well to being annotated. Adult annotators, at least, did not want to write on such poor quality paper, it was a last resort option. Further

---

[111] Acheson, *Early Modern English Marginalia*, 19.

[112] Acheson, *Early Modern English Marginalia*, 23.

extrapolating, those who did need to use chapbooks may have used pencils to avoid sinking, but which faded as exemplified by the pencil marks that *were* present in the model's detections. The most obvious reasoning for the lack of marginalia, thought, is that chapbooks which saw have use were likely damaged and disintegrated over time due to the poor materials used in their construction.

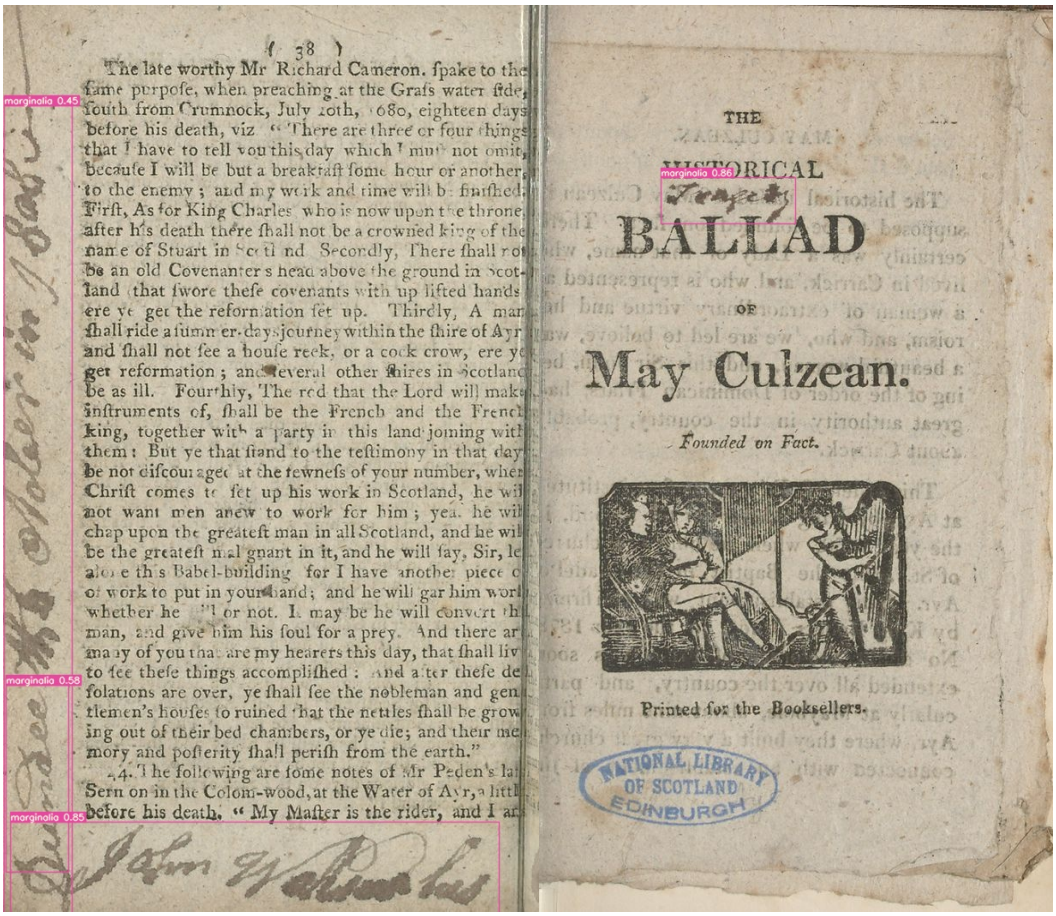


Figure 1010. **Left:** John Watson's absorbed ink. *The Life and Prophecies of Mr. Alexander Peden*, ca. 1800, National Library of Scotland, Associated metadata file from Data Foundry: 104186661-mets.xml. **Right:** Ink bleeding to the point of illegibility. Printed for the booksellers, *The Historical Ballad of May Culzean*, ca. 1817, National Library of Scotland, http://digital.nls.uk/104184173.

Conclusions

In this work, I have identified key issues which digitised cultural heritage collections and by extension, institutions, face during this time of rapid technological development, specifically in the realm of machine learning. I address these key issues through the development of an application to situate the data which machine learning models are built on in both their archival context and new context as metadata. I further demonstrate the potential positive impacts machine learning approaches might have for the study of history when used consciously with active effort made towards ethical usage.

The study of marginalia has evolved over time, with scholars exploring different aspects of this practice. Early works focused on using marginalia to reconstruct the lives and opinions of earlier scholars and expand on their published works. More recent studies have shifted towards analyzing the materiality of marginalia, examining how readers interacted with and used literature. Scholars have classified marginalia into different categories, including editing, interaction, and avoidance, and have explored the social, cultural, and intellectual contexts in which reading took place.

The presence of marginalia in physical and digital archives presents a challenge for researchers. Physical archives associated with academic and archival institutions have historically prioritized the preservation of pristine copies of books, often leading to the removal of marginalia assuming it was not composed by someone considered significant. Digitised archives, on the other hand, lack comprehensive metadata and search functionalities for marginalia. The lack of intentional curation of marginalia has limited its discovery and exploration. However, recent advances in computer vision and machine

learning, particularly object detection, offer new opportunities to identify and analyze marginalia within digital archives.

By examining object traces and using machine learning methods, researchers can expand their understanding of marginalia beyond the marginal doodles of a single reader. The approach I have developed allows us to explore many manuscripts at once, and look at questions of genre and categorization macroscopically, which in turn allow researchers to gain a deeper understanding of the historical, cultural, and social practices of early modern readers, as well as the ways in which texts were consumed, understood, and used in society. The automated detection of marginalia enables large and diverse corpora to be evaluated for the different forms and amounts of marginalia in a much faster way than locating it manually, providing researchers with more efficient methods to sift through works and focus on analysis. Overall, integrating machine learning into the study of marginalia presents exciting possibilities for advancing scholarship in this field.

It is clear from discussions surrounding digitised archives and machine learning methods that there is a need for more robust and comprehensive metadata. The current common practice of treating the content of digitised cultural heritage collections as transparent surrogates for physical objects overlooks the affordances and limitations of these digital artifacts. The use of machine learning in research further emphasizes the need for contextual information throughout the data lifecycle, from acquisition to analysis. The integration of archival materials into research also presents challenges in terms of handling diverse and abundant data, as well as the omission of archival metadata during the data collection process.

Understanding the foundations of machine learning models is essential for meaningful output. The biases and subjectivities inherent in the input data must be interrogated and contextualized to avoid reproducing harmful views or perpetuating discriminatory outcomes. When this is taken into consideration, descriptive metadata becomes a crucial part of holding the training data accountable, alongside providing transparency and accountability in the research process. Thus, by extension the development of tools and technologies that promote transparency, reproducibility, and accountability with a central focus on preserving the metadata created during the formation of training datasets is also crucial. The application I built provides a means to address these gaps and challenges in current digitised archives. By providing comprehensive metadata integrations through the annotation editor and metadata viewer, researchers are encouraged to understand and build their data through the viewing and creation of structured description, facilitating the meaningful annotation of data for computational research. Treating annotations as new digital objects and documenting their creation enhances familiarity with the training data and creates a more robust log of the data, ensuring transparency and accountability throughout the research process.

Essentially, the integration of robust and comprehensive metadata within digitised archives and machine learning workflows is essential for understanding the context, limitations, and biases of digital artifacts. By drawing on lessons from archival studies and implementing tools that prioritize transparency and accountability, researchers can make the most of digitised collections and ensure ethical and inclusive practices in data collection and analysis. The described application serves as an example of how these

principles can be applied in practice, facilitating the creation and annotation of training data while maintaining a clear record of the data's origins and transformations.

The use of machine learning models for the detection of marginalia in digitised chapbooks has both advantages and limitations. The application of machine learning in this context allowed for the quick and efficient identification of marginalia, providing a wealth of insights into how readers interacted with chapbooks in the past. It also allowed for the identification of patterns and trends in the annotations through marks of ownership, engagement with the text, and practical uses of the chapbook's surface as a place for quick calculations or writing practice. At the same time, there were still a significant number of false positives. These false positives included ink smudges, bleeding, and printing errors, which demonstrate the challenges of detecting marginalia in a dataset with complex and diverse page elements.

Nevertheless, the application of machine learning in the study of marginalia offers a new approach to understanding the materiality and readership of chapbooks. It provides a novel way to explore the annotations made by readers and the ways in which they engaged with the texts. This research demonstrates the viability and potential of both digitised collections as data and machine learning models in the study of cultural heritage collections.

As technology progresses, it is important to consider the ethical implications of using digitised collections as data for machine learning. Institutions holding these collections should actively participate in making their digitised collections open and accessible, while also implementing responsible data practices. If the institution is unable to distribute their data in multiple formats, there should be clear documentation on their

format of choice, how it is used, and ideally information or links to technical resources that allow transformation into other formats in a straightforward and reproducible way.[113] To harness the full potential of their content, cultural heritage institutions cannot only rely on the ability of the researchers to access their data through unmonitored and time consuming means such as webscraping. Instead, they must invest in more suitable ways to share their data, and in digital curation with a considerably broader scope of use, while also integrating their responsibilities to the content of their data regarding any ethical issues and inequities that may be present. Further, when the collections as data is applied to a project, guidelines, and checklists such as the "Collections as ML Data" checklist proposed by Benjamin Lee can help researchers and practitioners navigate the challenges and ethical considerations involved in using cultural heritage collections as data.

Overall, the use of machine learning models for the detection of marginalia in digitised chapbooks has the potential to enrich our understanding of the past and provide valuable insights into the materiality of these cultural artifacts. By engaging with these collections as data, researchers can shed light on the ways in which readers interacted with these texts, creating a more nuanced understanding of their historical, social, and cultural significance.

Future Directions

Several promising avenues for further exploration have emerged from this project. The simplest to explore would be to enhance the model's capabilities by expanding the training data. Incorporating marginalia that is on pages which are noisier and more

---

[113] Neudecker, "Cultural Heritage as Data," 3.

flawed such as those detected in the chapbooks collection would be an effective first step. This expansion would encompass a broader spectrum of paper quality, thereby improving the model's adaptability. Furthermore, including more instances of roughly printed text accompanied by marginalia can refine the model's ability to discern and differentiate between these elements accurately.

The accidental detection of thumb prints within the model's output gestures toward ways in which the method I utilize for training can be repurposed for investigating other more obscure facets of book history. Since the model has shown it is capable of finding trace marks even without intentional training, it would be interesting to train a model using the object traces found in early modern books which Smyth discusses.[114] This would, however, be much more of a challenging task as Smyth struggled to compile even just the few object traces he studied in his work. Gathering enough examples of object traces to create a dataset from and effectively train a model with would be difficult.

Extending the application of the object detection approach to other types of historical documents represents a natural progression. For instance, the technique could be adapted to unearth marginalia in digital collections such as EEBO. Additionally, the methodology holds potential in investigating more obscure facets of book history, such as object marks, inspired by Adam Smyth's exploration in "Object Traces in Early Modern

---

[114] Acheson, *Early Modern English Marginalia*, 51.

Books." It is acknowledged that constructing a suitable dataset for training in this area would be more challenging but equally rewarding.

When considering the results of this project in the context of digital cultural heritage collections themselves, there are multiple ways that machine learning models such as the one I produced can enhance these resources. First and foremost, there is immense potential for the enrichment of the current object metadata. Pages could be flagged as containing marginalia with consistency making it easier to search for and find with a cultural heritage institution's collections. This is not only beneficial for the researcher, but also for the institution, allowing for them to discover new readers present in their holdings and in turn, have a better understanding of their collections. As the Uppsala University's project points out, once the marginalia is detected, metadata could be even further enhanced through the automated transcription of the marginalia. Further, if the marginalia found across collections is substantial enough and patterns are identified, the marginalia could be coupled with an automated classifier system, enabling automated identification of marginalia patterns.

Closing Thoughts

Historical study in our digital age has been marked by the transformative impact of digitised resources, which have opened up new avenues for research and exploration, expanding the horizons of historical inquiry. This work has delved into the challenges and opportunities presented by the intersection of digitised cultural heritage collections and machine learning. The development of an application aimed at contextualizing training data for machine learning has showcased the potential positive impact of leveraging machine learning for historical analysis when accompanied by conscious

efforts toward ethical usage. The study of marginalia within digitised archives has evolved from a focus on individual readers to a broader exploration of social, cultural, and intellectual contexts. Machine learning methods, particularly object detection, have opened new avenues for understanding marginalia's role in history.

The first part of this work detailed the evolution of marginalia study, emphasizing the challenges faced by both physical and digital archives in preserving and providing access to these annotations. Then, the importance of robust metadata in the age of machine learning was highlighted, underlining the need for transparency, accountability, and ethical considerations throughout the research process as emphasized in the creation of my application. Finally, the advantages and limitations of using machine learning models to detect marginalia, and historical data broadly, were laid out, emphasizing the significance of ethical data practices in digitised collections. The future directions explored in this final section suggest avenues for expanding the model's capabilities, as well as investigating new facets of book history and digital archival practice as a result.

In conclusion, this work has contributed to bridging the gap between digital cultural heritage collections and machine learning, emphasizing the potential for meaningful insights while advocating for responsible and ethical practices. By integrating technology, humanist thought, and historical analysis, this research advances our understanding of the past and offers a roadmap for further exploration in this dynamic and evolving field.

Bibliography

"About the Collections in Calisphere." *Calisphere*. Accessed August 21, 2023. https://calisphere.org/overview/.

Acheson, Katherine, ed. *Early Modern English Marginalia*. New York: Routledge, 2019. https://doi.org/10.4324/9781315228815.

Adler, Noah, and Justin Hall. "Matt 28:19 - 28:20, Pg 141." *Manuscripts of Lichfiled Cathedral*. Accessed August 21, 2023. https://lichfield.ou.edu/file/14428.

"Archaeology of Reading," September 2014. https://archaeologyofreading.org/.

Axelsson, Adam, Liang Cheng, Jonas Frankemölle, and Ekta Vats. "Marginalia and Machine Learning: Handwritten Text Tecognition for Marginalia Collections." arXiv, March 2023. https://doi.org/10.48550/arXiv.2303.05929.

Baio, Andy. "Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator." *Waxy.org*, August 2022. https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. Virtual Event Canada: ACM, 2021. https://doi.org/10.1145/3442188.3445922.

Blodgett, Rachael. "Frequently Asked Questions NASA." Text. *NASA*, January 2018. http://www.nasa.gov/feature/frequently-asked-questions-0.

Bonde Thylstrup, Nanna, Daniela Agostinho, Annie Ring, Catherine D'Ignazio, and Kristin Veel, eds. *Uncertain Archives: Critical Keywords for Big Data*, 2021. https://doi.org/10.7551/mitpress/12236.001.0001.

Brayman Hackel, Heidi. *Reading Material in Early Modern England: Print, Gender, and Literacy*. Cambridge, U.K.; New York: Cambridge University Press, 2005.

Brousseau, Chantal. "Interrogating a National Narrative with GPT-2." *Programming Historian*, October 2022. https://doi.org/https://doi.org/10.46430/phen0104.

ChantalMB. "ChantalMB/Issap-Image-Annotator," February 2022. https://github.com/ChantalMB/issap-image-annotator.

Cordell, Ryan. ""Q i-Jtb the Raven": Taking Dirty OCR Seriously." *Book History* 20, no. 1 (2017): 188–225. https://doi.org/10.1353/bh.2017.0006.

Crymble, Adam. *Technology and the Historian: Transformations in the Digital Age*. Champaign, IL: University of Illinois Press, 2021. https://doi.org/10.5406/j.ctv1k03s73.

D'Ignazio, Catherine, and Lauren F. Klein. *Data Feminism*. Cambridge, MA: The MIT Press, 2020. https://doi.org/10.7551/mitpress/11805.001.0001.

Derrida, Jacques. *Of Grammatology*. Translated by Gayatri Chakravorty Spivak. Corrected ed. Baltimore: Johns Hopkins University Press, 1997.

Dutta, Abhishek, and Andrew Zisserman. "The VIA Annotation Software for Images, Audio and Video." In *Proceedings of the 27th ACM International Conference on Multimedia*, 2276–79. MM '19. New York, NY, USA: Association for Computing Machinery, 2019. https://doi.org/10.1145/3343031.3350535.

"Energy Use in Sweden." *Sweden.se*, November 2022. https://sweden.se/climate/sustainability/energy-use-in-sweden.

Fleming, Juliet. *Cultural Graphology: Writing After Derrida*. University of Chicago Press, 2016.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. "Datasheets for Datasets," December 2021. https://doi.org/10.48550/arXiv.1803.09010.

"Getty API Documentation." *Getty*. Accessed August 18, 2023. https://data.getty.edu/museum/collection/docs/#attribution.

Graham, Shawn, and Justin Walsh. "Recording Archaeological Data from Space." *International Space Station Archaeological Project*, February 2022. https://issarchaeology.org/how-do-you-get-from-an-astronauts-photo-to-usable-archaeological-data/.

Grindley, Carl James. "Reading Piers Plowman C-Text Annotations: Notes Toward the Classification of Printed and Written Marginalia in Texts from the British Isles 1300-1641." In *The Medieval Professional Reader at Work: Evidence from Manuscripts of Chaucer, Langland, Kempe, and Gower*, edited by Kathryn Kerby-Fulton and Maidie Hilmo, 73–141. Victoria, BC: English Literary Studies, 2001.

Harvey, Gabriel, and George Charles Moore Smith. *Gabriel Harvey's Marginalia*. Stratford-upon-Avon: Shakespeare Head Press, 1913.

Jardine, Lisa, and Anthony Grafton. ""Studied for Action": How Gabriel Harvey Read His Livy." *Past & Present*, no. 129 (1990): 30–78. https://www.jstor.org/stable/650933.

Jo, Eun Seo, and Timnit Gebru. "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 306–16. FAT* '20. New York, NY, USA: Association for Computing Machinery, 2020. https://doi.org/10.1145/3351095.3372829.

Lacoste, Alexandre, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. "Quantifying the Carbon Emissions of Machine Learning." *arXiv Preprint arXiv:1910.09700*, 2019.

"Laion-Aesthetic-6pls: Images 1582553." *LAION-Aesthetics V2 6+*. Accessed August 21, 2023. http://laion-aesthetic.datasette.io/laion-aesthetic-6pls/images/1582553.

"Lauriston Castle Collection." Accessed August 16, 2023. https://digital.nls.uk/catalogues/special-and-named-printed-collections/?id=598.

Lee, Benjamin. "Compounded Mediation: A Data Archaeology of the Newspaper Navigator Dataset." *Digital Humanities Quarterly* 015, no. 4 (December 2021). http://www.digitalhumanities.org/dhq/vol/15/4/000578/000578.html.

Lee, Benjamin Charles Germain. "The 'Collections as ML Data' Checklist for Machine Learning and Cultural Heritage." *Journal of the Association for Information Science and Technology* n/a, no. n/a (May 2023). https://doi.org/10.1002/asi.24765.

Lerer, Seth. "Devotion and Defacement: Reading Children's Marginalia." *Representations* 118, no. 1 (2012): 126–53. https://doi.org/10.1525/rep.2012.118.1.126.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. "Microsoft COCO: Common Objects in Context." arXiv, February 2015. https://doi.org/10.48550/arXiv.1405.0312.

Mak, Bonnie. "Archaeology of a Digitization." *Journal of the Association for Information Science and Technology* 65, no. 8 (2014): 1515–26. https://doi.org/10.1002/asi.23061.

Manley, K. A. "Scottish Circulating and Subscription Libraries as Community Libraries." *Library History* 19 (July 2013). https://doi.org/10.1179/lib.2003.19.3.185.

Milligan, Ian. "We Are All Digital Now: Digital Photography and the Reshaping of Historical Practice." *The Canadian Historical Review* 101, no. 4 (2020): 602–21. https://doi.org/https://doi.org/10.3138/chr-2020-0023.

Mordell, Devon. "Critical Questions for Archives as (Big) Data." *Archivaria* 87 (2019): 140–61. https://proxy.library.carleton.ca/login?qurl=https%3A%2F%2Fwww.proquest.com%2Fscholarly-journals%2Fcritical-questions-archives-as-big-data%2Fdocview%2F2518871266%2Fse-2%3Faccountid%3D9894.

Moss, Michael, David Thomas, and Tim Gollins. "The Reconfiguration of the Archive as Data to Be Mined." *Archivaria*, November 2018, 118–51. https://archivaria.ca/index.php/archivaria/article/view/13646.

Neudecker, Clemens. “Cultural Heritage as Data: Digital Curation and Artificial Intelligence in Libraries.” In *Proceedings of the Third Conference on Digital Curation Technologies (Qurator 2022), Berlin, Germany, Sept. 19th-23rd, 2022*, edited by Adrian Paschke, Georg Rehm, Clemens Neudecker, and Lydia Pintscher, Vol. 3234. CEUR Workshop Proceedings. CEUR-WS.org, 2022. https://ceur-ws.org/Vol-3234/paper2.pdf.

“Omniscribe.” BuildUCLA, February 2019. https://github.com/collectionslab/Omniscribe.

Orgel, Stephen. *The Reader in the Book: A Study of Spaces and Traces*. Oxford, UK: Oxford University Press, Incorporated, 2015. http://ebookcentral.proquest.com/lib/oculcarleton-ebooks/detail.action?docID=4310757.

Oxford English Dictionary. “Marginalia, n., Etymology.” Oxford University Press, 2023. https://doi.org/10.1093/OED/7050641376.

Padilla, Thomas. “On a Collections as Data Imperative,” 2017. https://escholarship.org/uc/item/9881c8sv.

Palmer, Philip. “Annotated Books at UCLA: Wider Applications of the AoR Schema Archaeology of Reading.” *Archaeology of Reading* , September 2018. https://archaeologyofreading.org/annotated-books-at-ucla-wider-applications-of-the-aor-schema/.

“Results – Search Objects – eMuseum.” *Peabody Museum of Archaeology & Ethnology*. Accessed August 21, 2023. https://collections.peabody.harvard.edu/search/daguerreotype/objects.

Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. “High-Resolution Image Synthesis with Latent Diffusion Models,” 2021. https://arxiv.org/abs/2112.10752.

Schramowski, Patrick, Manuel Brack, Björn Deiseroth, and Kristian Kersting. “Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models.” In *Proceedings of the 22nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC: arXiv, 2023. https://doi.org/10.48550/arXiv.2211.05105.

Scotland, National Library of. “Chapbooks Printed in Scotland.” National Library of Scotland. Accessed May 30, 2023. https://doi.org/10.34812/VB2S-9G58.

Scott-Warren, Jason. “Reading Graffiti in the Early Modern Book.” *The Huntington Library Quarterly*, 2010. https://www.proquest.com/docview/763492186/abstract/832093B897F1447APQ/1.

Sherman, William H. *Used Books: Marking Readers in Renaissance England*. Philadelphia, PA: University of Pennsylvania Press, 2008. https://www.jstor.org/stable/j.ctt3fhgzw.

"SvelteKit • Web Development, Streamlined." Accessed August 21, 2023. https://kit.svelte.dev/.

team, The MicroPasts. "Crowdfuelled and Crowdsourced Archaeological Data." *MicroPasts: Crowd Sourcing Platform*. Accessed August 21, 2023. https://crowdsourced.micropasts.org/.

"UCLA / William Andrews Clark Memorial Library." *Calisphere*. Accessed August 21, 2023. https://calisphere.org/institution/62/collections/.

Victoria, and Albert Museum. "Victoria and Albert Museum Collections Data," 2021. https://collections.vam.ac.uk/.

Wang, Chien-Yao, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors." *arXiv Preprint arXiv:2207.02696*, 2022. https://doi.org/10.48550/arXiv.2207.02696.

Weaver, Edmund. *Wigmaker's Account Book.* London: A. Parker for the Company of Stationers, 1737. https://calisphere.org/item/ark:/21198/n14s4d/.

Whitaker, Elaine E. "A Collaboration of Readers: Categorization of the Annotations in Copies of Caxton's Royal Book." *Text* 7 (1994): 233–42. https://www.jstor.org/stable/30227702.

Yale, Elizabeth. "The History of Archives: The State of the Discipline." *Book History* 18 (2015): 332–59. https://doi.org/10.1353/bh.2015.0007.

Ziegler, S. L. "Open Data in Cultural Heritage Institutions: Can We Be Better Than Data Brokers?" *Digital Humanities Quarterly* 014, no. 2 (June 2020). http://www.digitalhumanities.org/dhq/vol/14/2/000462/000462.html.

Appendix A: Accessing Digital Components

- The digital version of this essay is hosted here: https://chantalmb.github.io/MRE-MitM-2023/
- The GitHub repository for the application is located at: https://github.com/ChantalMB/MRE-RocketAnno
    - This contains the application downloads, as well as all of the application's code.
    - The README file lists current limitations of the application, as well as ways I wish to improve it in the future.
- The GitHub repository for the case study is located at: https://github.com/ChantalMB/MRE-Detecting-Marginalia
    - This repository contains the code used for webscraping and training the machine learning model, as well as the model itself.
    - It also contains the project save file from the application which was produced when annotating images for this case study.

Appendix B: The Full "Collections as ML Data" Checklist

The cultural heritage collection as data

Here, a distinction is drawn between the cultural heritage collection being studied and the training dataset being utilized for the machine learning model. For example, a project might utilize a pre-trained model to generate embeddings for a photo collection. In this section, we consider the cultural heritage collection itself; in the section "The Machine Learning Model," we consider the machine learning model's training data.

1. Dataset composition:
    a. Who or what is depicted in the dataset? (Gebru et al., 2020)
    b. If the dataset depicts people, are any specific subgroups of people represented? Are any specific individuals personally identifiable? (Gebru et al., 2020)
    c. If the dataset depicts people, are any individuals still living? Does this project comply with privacy laws in countries where it will be shared?
    d. What medium is the dataset? (image, video, text, web archive, etc.)
    e. How large is the dataset, both in cardinality and in disk storage?
    f. What metadata is available for the dataset items? (Holland et al., 2018)
    g. Does copyright impact this dataset? If so, how? (Cordell, 2020; Gebru et al., 2020; Jakeway et al., 2020; Padilla, 2018)
    h. Does this dataset pertain to a difficult history? If so, what extra precautions are being taken?
2. Collecting process and curation rationale (language borrowed from Bender and Friedman (**2018**)):
    a. Who curated the cultural heritage collection from which this dataset is derived?
    b. What organization or institution was the collection created for?
    c. What funding was utilized (if known)?
    d. What collection process was utilized? (Bender & Friedman, 2018)
    e. When was the collection assembled? (i.e., when were the photographs taken or ethnographies recorded?)
    f. What instruments were utilized to create the collection? (i.e., a recording device, camera, etc.)
    g. If people are included, did individuals consent at the time of collection?
    h. What were the decision-making processes behind the collection's curation? (Bender & Friedman, 2018)
    i. What is unknown about the collection process and curation rationale?
3. Digitization pipeline (only applicable if the dataset is a digitized version of a physical collection):
    a. Who selected what was digitized?
    b. What organization or institution oversaw the digitization?
    c. What funding was utilized?

    d. What criteria were utilized for determining what was digitized? (Cordell, 2020)
    e. What were the steps in the digitization pipeline? (For example, in the case of photos, what scanners were used to digitize the documents? In the case of documents, what OCR engines were utilized?)
    f. What metadata was algorithmically produced?
4. Data provenance:

    a. What is the provenance of the dataset, from collection through digitization? (Bender & Friedman, 2018; Diakapoulos et al., n.d.; Holland et al., 2018)
    b. Is any part of the provenance unknown?
5. Crowd labor:

    a. Have volunteers or crowd workers added metadata to the dataset? (Cordell, 2020; Jakeway et al., 2020; Padilla, 2018)
    b. If so, how were they recruited and compensated?
    c. If so, what metadata did they produce? (i.e., transcriptions, annotations, etc.)
6. Additional modification:

    a. Were any additional steps taken after collection curation and digitization in order to produce the dataset in question? (i.e., Were any items removed? Were any additional metadata added? etc.)

The machine learning model

Note: if multiple machine learning models were utilized in the project, this step should be completed for each model.

1. Overview:

    a. What model architecture has been utilized? (Mitchell et al., 2019)
    b. What is the task that the model is being deployed to perform?
    c. Who trained, finetuned, and/or deployed this model? (Mitchell et al., 2019)
    d. Across what organizations or institutions did this training, finetuning, and/or deployment take place? (Mitchell et al., 2019)
    e. What funding was utilized? (Gebru et al., 2020)
2. Training/finetuning:

    a. Was the model trained from scratch?
    b. If so, what data was used to train the model? (Mitchell et al., 2019)
    c. If not, was a pre-trained model utilized? Where can more information on the pre-trained model be found? (Mitchell et al., 2019)
    d. Was the pre-trained model finetuned? If so, what data was utilized for finetuning?

    e. If training or finetuning was performed, what computational resources were utilized?
3. Evaluation:

    a. How was the model's performance evaluated? (Mitchell et al., 2019)
    b. What data was used for evaluation? (M. Arnold, Bellamy, et al. (2019); Mitchell et al., 2019)
    c. If the model involves data pertaining to people, has the model been audited for fairness and bias using tools such as FairLearn? (M. Arnold, Bellamy, et al. (2019); Bird et al., 2020; Diakapoulos et al., n.d.; Jakeway et al., 2020; Madaio et al. (2020); Reisman et al., 2018)
    d. Have any tools been utilized to generate explanations for predictions (i.e., LIME Ribeiro et al. (2016), SHAP Lundberg and Lee (2017), TCAV (Kim et al., 2018)) and modify the model in response? (M. Arnold, Bellamy, et al. (2019); Cordell, 2020; Diakapoulos et al., n.d.; Padilla, 2020; Ribeiro et al., 2020)
4. Deployment:

    a. How was the model deployed? Was it used to make a single pass over the cultural heritage dataset in question, or will it be continuously deployed?
    b. What computational resources were utilized for deployment?
    c. Are the metadata generated by the machine learning model (embeddings, classifications, etc.) available as project deliverables?
5. Release:

    a. Has the resulting model been made available for download? (*if no, the following questions can be skipped*)
    b. What license has been provided? (Mitchell et al., 2019)
    c. Who are the primary intended users, and what are the intended use cases? (Mitchell et al., 2019)
    d. Does this model have applicability outside of cultural heritage collections?
    e. What are ways that this model could be misused, either intentionally or unintentionally? (Madaio et al., 2020; Mitchell et al., 2019)
6. Environmental impact:

    a. What were the carbon emissions produced by training, finetuning, and/or deploying this model? (Cordell, 2020; Lacoste et al., 2019; Strubell et al., 2019)
    b. How does the environmental impact of this model compare to that of other components of the project, such as a collection's digitization or stakeholders' flights to relevant conferences?

Organizational considerations

1. Stakeholders:

    a. What stakeholder groups are involved in this project? (Cordell, 2020)

   b. What is each project member's familiarity with machine learning? (Cordell, 2020; Jakeway et al., 2020)
   c. What is each project member's familiarity with cultural heritage collections as data?
   d. Has the project notified and sought input from all potentially relevant stakeholder groups, such as those included within the cultural heritage dataset itself? (Madaio et al., 2020; Reisman et al., 2018)
   e. Do groups affected by the project, such as individuals and communities directly represented within the cultural heritage dataset, have an avenue for contacting project staff and seeking recourse? If so, whom should they contact? If not, why not? (Diakapoulos et al., n.d.; Mitchell et al., 2019; Reisman et al., 2018)
2. Use of machine learning:

   a. Was it necessary to use machine learning for this project?
   b. If so, why?
   c. If not, why was machine learning still utilized?
   d. What are potential critiques of applying machine learning in this context?
3. Organizational context:

   a. Can this project be used to build data fluency within the organization or institution? (Padilla, 2020)
   b. Do there exist programs or paths for training staff affiliated with the project to develop machine learning skillsets? (Cordell, 2020; Padilla, 2020)
   c. Do there exist programs or paths for training staff affiliated with the project to develop fluency with cultural heritage collections?
4. Project deployment and launch:

   a. Who is the target audience of this project? (Madaio et al., 2020)
   b. How does the target audience align with the audiences that the institution or organization is hoping to engage?
   c. If the target audience of the project is the public, does it make an attempt to educate the public regarding the machine learning approaches employed?
   d. *Did the project launch reach the intended audience?**
   e. *Has the project received feedback from stakeholders, including the audience? If so, what feedback has been received?**
   f. *Has the launch of the project resulted in any changes to the project?**

   (* = *to be completed post-launch*)

Copyright, transparency, documentation, maintenance, and privacy

1. Copyright:

    a. Building on question 1.1.g, does copyright impact the dataset, model, code, or deliverables for the project? (Cordell, 2020; Gebru et al., 2020; Jakeway et al., 2020; Mitchell et al., 2019; Padilla, 2018)
    b. If they are made available, what licenses have been chosen?
    c. If they are proprietary, how does this impact re-use?
2. Transparency and re-use:

    a. Can the project be audited by outsiders? If so, is there funding available to support outside audits? (Mitchell et al., 2019; Reisman et al., 2018)
    b. Is the code created for the project extensible for other cultural heritage researchers? (Padilla, 2020)
    c. If so, does the project provide any tutorials or toolkits for re-use?
3. Documentation:

    a. Does the project have documentation? (Katell et al., 2019)
    b. If so, is the documentation interpretable by the project's audience?
    c. Is the project reproducible to an outside researcher, given the documentation available?
4. Privacy:

    a. If the project is hosted online, are data on visitors collected? If so, what kinds of user data are collected? (Cordell, 2020)
    b. Is visitor consent gained before gathering online data? (Cordell, 2020)
5. Maintenance:

    a. Will the project and code be maintained? (Gebru et al., 2020)
    b. If so, how frequently, and who will be responsible for maintaining it?

Appendix C: Technical Overview

Application Development

When using digitised cultural heritage collections as data, each step of the process holds the possibility of introducing unspoken assumptions or hidden transformations of the data. Thus, the digital historian must write with both reproducibility and transparency in mind, so that the reader can verify and trust the results and conclusions. Even historians engaging in methods they would not consider digital make similar transformations as they convert historical information into their notes and writing, although these transformations are not nearly so apparent; as Ian Milligan stated in his piece documenting the research practices of historians in the archive during this period of technological shifts, "we are all digital historians now."[115] Thus, before delving into the details of its creation, it is important to briefly consider the technical foundation of the application in order to understand the very first considerations that went into how the end product took form.

The code for this application was written using SvelteKit, a framework for building web applications using a specialised implementation of JavaScript.[116] When the website is compiled, the step in the web development process where the code is converted into what is displayed on a web page, the code is converted into highly efficient vanilla JavaScript resulting in faster performance compared to traditional

---

[115] Ian Milligan, "We Are All Digital Now: Digital Photography and the Reshaping of Historical Practice," *The Canadian Historical Review* 101, no. 4 (2020): 620, https://doi.org/https://doi.org/10.3138/chr-2020-0023.

[116] "SvelteKit • Web Development, Streamlined," accessed August 21, 2023, https://kit.svelte.dev/.

frameworks that compile into multifaceted layers. It is this performance advantage, as well as ease of use, that led me to select SvelteKit for this task, since as an image annotator, users would be uploading and interacting with files, as well as populating files with data which can be computationally heavy tasks, thus having a performant framework was a must.

To transform the SvelteKit web application into a desktop application, I used Electron, an open-source software framework that allows developers to build cross-platform desktop applications using web technologies such as HTML, CSS, and JavaScript. Electron powers many popular desktop applications such as Discord, Slack, Notion, and even the application which I am writing in right now, Visual Studio Code. It functions by running the web application code using a Chromium browser engine, essentially turning the code into a browser itself designed to do the singular task which the web application's code instructs it to do. Due to its popularity, Electron has extensive documentation and a large community making it a good choice for developers such as myself who have little experience creating desktop-based software. The most significant limitation Electron presents is that by using web technologies to build desktop applications, there can be a slightly higher memory consumption and application size compared to "native" applications, which are software that is developed specifically for a particular operating system or platform and thus very able to be more optimized since they are developed with very specific hardware parameters in mind. Alternatives to Electron which focus on reducing memory consumption and application size have begun to be developed in recent years, the most notable being Tauri, however they optimize the application through using whatever browser engine the operating system comes pre-

installed with rather than installing a dedicated one; for example, if a user is opening the application on an Apple device, it will run using a Safari browser engine. Each browser has different development standards and ways which they display information, so if the application uses the browser engine based on the user's device, the developer ultimately has little control over how the application appears and functions outside of the operating system that they use to develop the application, unless they have access to multiple machines to perform testing, as well as the resources to tailor a version of the application to each browser engine.

I chose to make my application a desktop tool rather than publish it as a website to ensure both ease of collaboration and use. The application functions by creating and updating a project save file, which is simply a JSON file that contains all data surrounding the images selected for annotation, as well as any annotations drawn upon each image. JSON files, being a form of structured data, are both human readable and easy for a machine to manipulate in a consistent way. Additionally, they are small in size which makes them easily shareable. These factors combined have made the JSON file format popular which has resulted in many tools that can make them useable even outside of the application. When used in a project with collaborative annotation needs, the project save file can be placed in a shared code repository such as GitHub and versions can be managed using Git. This method also adds a level to transparency to a project using my application, as each step of annotating images and changes are being recorded through each push to the repository, assuming the repository is public. As a security measure, web browsers are not allowed access to a user's file system; when uploading a file to a website, a temporary fake path to the selected file is generated, and this is either used to

make a copy of the file which is then stored on the website's server (for example, Google Drive), or temporarily stored then discarded once the web page is closed, the latter of these options being very resource intensive if uploading a large number of files. Electron applications allow access to the file system, since although it makes use of a browser engine, this engine is installed and run locally on the user's device rather than being connected to the World Wide Web. When starting a new or existing project with my application, the user is first prompted to select the folder of images which they want to annotate. This establishes a path to where the images are on the user's computer since this does not get saved in the project save file, as paths to where the images are located are unique to each device. If the user wants to open an existing project, they will also be prompted to select the project save JSON file. The path to this file will also be saved so that when the user manually saves their project, this same file will be updated. Access to the file system also allows for the application to autosave the project save file, so should anything go wrong, the user will lose at most ten minutes of work. In summary, Electron ensures that the functional, visual, and file-based user experience is universal when using this tool.

Interface and Tooling

As a whole, when designing the interface of my application, I sent through the design around the user experience/design concepts of mapping, and the principle of familiarity. The principle of familiarity is concerned with the ability of an interactive system to allow a user to map prior experiences, either real world or gained from interaction with other systems, onto the features of a new system. By extension, mapping in this context is using a familiar imagery to invoke the action/operation which an

interactive element will perform. The layout of the app is similar to other popular tools for image manipulation, such as those within the Adobe Suite, MS Paint, or Windows Photo Viewer.

To the left of the window is a simple tool bar, which allows user to select the shape they want to use for annotating the image (set to a rectangle by default), perform basic manipulations like zooming in and out in a controlled manner, and "reset" the image to its original position. This tool bar is hovering over the largest component of the application, the image viewer and annotation canvas. This viewer utilizes technology that those in the humanities are likely already familiar with, even if they may not be aware of it. The image viewer itself uses OpenSeadragon, a tool for viewing high-resolution zoomable images, which is the technology largely behind image viewers used by digital archives. aside from using the tool bar's buttons, OpenSeadragon also allows user to manipulate the image using trackpad gestures as well as click-and-drag to move around the image. Annotorious works with OpenSeadragon to allow for annotations to be drawn on images viewed in the OpenSeadragon window; this combination has been leveraged for cultural heritage purposes before, one notable example being the Arts and Humanities Research Council crowdsourcing platform, MicroPasts, which allows for the public to assist with large scale archaeology, history and heritage tasks.[117] At the bottom right of the image viewer are arrows the user may use to switch from one image to the next, however they may also do so by using the left and right arrow keys.

---

[117] The MicroPasts team, "Crowdfuelled and Crowdsourced Archaeological Data," *MicroPasts: Crowd Sourcing Platform*, accessed August 21, 2023, https://crowdsourced.micropasts.org/.

Occupying the right side of the application window is the primary space for application and file management. At the top of this space, users can either return to the home menu should they want to begin a new project, choose to manually save their project, or select where they want the project save file to be saved. Below this is a file viewer, where users can add images they want to annotate from the image folder they selected when creating the project or remove them, as well as view a list of image files they have uploaded. The user can also jump to any of the images by clicking the relevant file name in the file list. Below the file list is a drop-down menu where the user may apply a filter that indicates whether images have or have not been annotated, which functions by highlighting the entry in the file list that matches the filter criteria. The search bar functions in the same highlighting manner, except it highlights the images which have metadata that matches the search term. Following this, there is an "Export" menu in which users can choose to export their annotated images into a variety of popular data formats used for training object detection models. The last section included in this side bar is a quick guide to how the application functions as a reminder to users who have just begun using the application or are returning to the project after a period time away from it.

Annotation Process and Dataset Formation

With the images for collections now downloaded and formatted, the process of annotating the images for use as training data was able to begin. I created a new folder for the project to contain the images I desired to annotate and the save file, and then into this folder I copied the 43 known images of chapbooks along with an assortment of randomly selected pages from the AoR and Early Modern Annotated Books collections. In the

application, I created a new project and in the annotation editor, I added the column "notes". Given the limited time frame I had available to dedicate to annotation, I felt that having one general purpose column would be sufficient to compliment the content of the default columns, allowing me space to make note of uncertainties or points of interest. I then added the images I wanted to annotate from the project folder into the workspace and began creating annotations. I ultimately ended up using the notes column to describe marginalia, particularly symbols, that I was unsure the meaning of. Once I finished annotating, I returned to these marginalia through using the application's keyword search to find the relevant pages and was able to research and clarify the meaning of these mysteries, such as the manicule, which I was unaware was once a commonly hand-drawn symbol to denote an important passage of text to the reader among other similar functions.[118]

Following annotation, I used the application's export functionality to save my images and their annotations in "You Only Look Once" (YOLO) format, since I intended to use a YOLO object detection model which will be discussed with more depth in the following section. In total, within the time frame I allotted to this task, 353 images of early modern book pages were annotated. Added into this set of images was 20 negative examples, that is, pages which contain no marginalia. This set of images was then randomly split using Python's machine learning tool library sklearn into an 85/15 training/test ratio, with 317 (~85%) images being used to *train* the model, 39 (~10%) images being used to *validate* the model while it trains, and 17 (~5%) images to test the

---

[118] Sherman, *Used Books*, 29.

results of the trained model. A standard recommendation for training an object detection model is to have 1500 images per new class, so to give this small dataset a fighting chance at successfully detecting marginalia, I applied various transformations using a modified version of a data augmentation command line tool to the original training images to create augmented versions of this data. These transformations modified the images in ways that preserve their essential features but introduce variation that can enhance the model's ability to generalize. Not only does data augmentation increase the amount of training data by generating multiple versions of each original image, but also by presenting the model with different variations of the same image (e.g., different rotations, flips, translations, zoom levels, etc.), data augmentation helps the model learn to recognize important patterns and features that are invariant to those transformations. At the same time, it also helps prevent the model from memorizing specific details of the training data which results in "overfitting"– we want a model that detects meaningful and relevant features rather than memorizing specific examples. In the context of this case study, each of the 317 training images were augmented through added noise, randomly changed contrast, randomly changed brightness, and randomly changed saturation. This expanded the total training image set to 1585 (317*4), however, even then, within the greater context of machine learning, this dataset is very small at only 795 MB.

The Machine Learning Model

For the purpose of object detection, I chose to use the YOLOv7 model.[119] In general, the YOLO family of models function by processing the entire image just once to detect objects instead of iteratively analyzing an image multiple times at different scales to identify objects as other object detection models have traditionally done, which makes YOLO models perform the task of object detection faster and more efficiently compared to other models. YOLO models do this by first dividing a picture into a grid of smaller sections, and then for each of these smaller sections, the model tries to predict whether or not there is an object present in that grid, and if so, what kind of object it might be through drawing bounding boxes around the entirety of the possible object and assigning it a label. For each of these boxes drawn, the model will also assign a confidence score, which indicates how sure the model is that an object in that box. If the score is high, it means the model is quite confident, and conversely, if it is low, it is less certain. Lastly, after predicting objects in all of the image's sections, the model eliminates any object that may have multiple bounding boxes drawn around it using a technique called Non-Maximum Suppression, which selects the best bounding boxes by keeping those with the highest confidence score and removing overlap or redundancy.

Compared to models that break the image down into more granular, pixel-level segmentation, YOLO models' single-pass approach can struggle with smaller objects or complex scenery, however, should these limitations be considered when creating the

---

[119] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors," *arXiv Preprint arXiv:2207.02696*, 2022, https://doi.org/10.48550/arXiv.2207.02696.

training dataset, these issues can be countered through methods such as the addition of negative examples or data augmentation as described previously. Another advantage of the YOLO algorithm is that there is typically little need for negative examples; in the initial step of dividing the inputted image into a grid then analyzing each segment, the model will naturally learn how to identify “no object” since a majority of images will not contain objects in every grid segment. However, when the images being used to train the model are “busy” with the objects being detected possibly being present across the entire page, such as with images of traffic or in our case, with pages of text and prolific annotation, training results can be improved through the intentional addition of negative examples to reduce false positives caused by heavy overlap of desired objects and background objects.

The goal of this model is to detect marginalia present in the NLS chapbooks dataset in order gain a sense of how chapbooks were used by those who interacted with them. It is unlikely that the model will achieve exceptionally high performance due to the small amount of training data, however with YOLOv7’s efficient learning capabilities it is likely that the model will achieve a standard of performance that will be acceptable to garner the “bird’s eye view” of the chapbooks desired for this experiment. The alternative to machine learning for this project would be algorithmic computer vision, using a series of image augmentations such as feature extraction or template matching to identify areas of marginalia on a page. However, these techniques would not be able to achieve the same level of performance and adaptability as modern machine learning approaches, especially for complex and large-scale object detection tasks such as the one at hand. Machine learning techniques are much stronger at generalization, adapting better to

various challenges, such as object variations, scale changes, occlusions, and complex scene contexts.

My dataset is small, so rather than training a model from scratch which would likely yield poor results due to this limitation, I chose to use transfer learning with one of YOLOv7's pretrained models. Transfer learning is a machine learning technique where a model that has been trained on one task is repurposed or "finetuned" for a different but related task. Instead of starting from scratch, the knowledge gained from solving one problem is transferred to help solve a different problem; this not only saves time and resources since the model is not being built from the ground up, it also builds upon what the model has already learned about recognizing characteristics such as shapes and patterns, making it more efficient and effective at expanding this palette. Under the guidance of the YOLOv7 paper, while also accounting for the technical parameters of my dataset and hardware being used, I chose the YOLOv7-E6 pretrained model to build off of. The YOLOv7-E6 model is designed for larger input sizes, that is to say, higher resolution images, which helps in capturing smaller details in an image, as well as for use with cloud GPUs.[120] Cloud GPUs are remote graphics processing units that can be rented or accessed on-demand through cloud computing services for various computational tasks, particularly those involving machine learning workloads since machine learning makes use of the parallel processing capabilities GPUs possess which allow for multiple calculations to be done simultaneously, making complex computation faster and more efficient. In general, cloud GPUs are used because they allow for the use of a more

---

[120] Wang, Bochkovskiy, and Liao, "YOLOv7," 6.

powerful GPU than one typically has on hand, however they can also be a more environmentally conscious choice; cloud computing allows for resource sharing among multiple users which reduces the overall energy consumption compared to individuals running their own hardware.

To train my model, I chose to use an NVIDIA RTX A6000 GPU from provider Vast.ai, a market-based cloud computing platform which allows all compute providers large and small to easily share their devices' spare capacity. In using the spare capacity of an already running device, platform such as Vast.ai help democratise advanced machine learning research by those without institutional access to compute through more affordable pricing than other cloud computing providers who own and allocate dedicated GPUs to users as requested. I chose to use an an RTX A6000 due to the high amount of virtual memory which is necessary when training with a higher image resolution, but also because it was available from a Swedish data centre; nearly 75% of electricity production in Sweden comes from renewable, green energy sources, meaning that using a GPU located in Sweden will likely produce less carbon dioxide emissions compared to a GPU present elsewhere.[121] Undoubtedly the most resource intensive component of this project was training the model. Although I did not track the carbon emissions of the model as it trained, I utilized the Machine Learning CO2 Impact calculator, which uses the formula of power consumption x time x carbon produced based on the local power grid to

[121] "Energy Use in Sweden," *Sweden.se*, November 2022, https://sweden.se/climate/sustainability/energy-use-in-sweden.

estimate the carbon emissions of my training.[122] Since Vast.ai follows a market place shared compute model, I used the Stockholm-based Amazon Web Services datacenter to stand in for the Swedish datacenter which I rented the RTX A6000 used for training from. Approximately 15 hours of computation was performed using this GPU, which has a TDP of 300W, between initial attempt at transfer learning and a following attempt after updating hyperparameters for better performance. Region eu-north-1 has a carbon efficiency of 0.05 $kgCO_2eq/kWh$ thus the total emissions are estimated to be 0.23 $kgCO_2eq$, which is the equivalent of around 1km driven in a car with an internal combustion engine.

To train my model, I was able to follow the command for transfer learning provided in the YOLOv7 GitHub repository. There were a small number of changes to the training code suggested in the repository's issues page which I applied prior to starting training; firstly, I lowered the learning rate hyperparameter. A hyperparameter in the context of machine learning is essentially the setting of the model; often, the defaults provided work fine, however, results sometimes can be improved through slight modifications. The learning rate is like a step size that determines how quickly or slowly a model adjusts to its new parameters as it learns from training; a lower learning rate means taking smaller, more "detailed" steps towards learning which in turn improves how well the model is able to learn. Secondly, I updated the ar_thr variable in dataset.py from 20 to 100; ar_thr refers to the aspect ratio threshold which is a value that determines

---

[122] Alexandre Lacoste et al., "Quantifying the Carbon Emissions of Machine Learning," *arXiv Preprint arXiv:1910.09700*, 2019.

whether an object's bounding box should be included in the training data or not based on its aspect ratio. Since I recalled from the process of annotation that some bounding boxes were quite tall or thin, I chose to make the accepted aspect ratio larger to ensure that no bounding boxes were eliminated from the images while training.

The key metrics produced upon testing a YOLOv7 model to evaluate its performance are precision, recall, and mean average precision (mAP). Precision is the ratio between actual positive detections and all positive detections; in the context of this model, that would be the measure of marginalia detected on the page out of all the marginalia actually present. Recall indicates how well a model correctly detects the objects broadly; thus, for all the marginalia present, recall tells us how many were correctly detected. The mAP compares the bounding box that was drawn by the annotator, the ground-truth bounding box, to the bounding box detected by the model and returns a score; this score determines if an object has been successfully detected or not. YOLOv7 evaluates the model using mAP@.5 specifically; the appended .5 indicates the Intersection-over-Union (IoU) threshold, which measures the minimum overlap between the model's predicted boundary and the ground truth for the detection to be considered correct. After training the model for 150 epochs, a single epoch being one pass through the entire training dataset during the training process, when tested on the test set of pages the model outputted a precision score of 77.6%, a recall score of 79.5%, and a mAP@.5 of 77.1%. So, when the model predicts marginalia, it is generally correct, and the model is able to find and capture a substantial portion of the actual marginalia. Considering the small size of the training dataset and the goals of this project as a whole, these results are

acceptable enough that the model should provide reasonably accurate detections of marginalia when applied to the NLS chapbook dataset.