DHQ: Digital Humanities Quarterly
Changing the Center of Gravity: Transforming Classical Studies Through Cyberinfrastructure
2009
Volume 3 Number 1

Digitizing Latin Incunabula: Challenges, Methods, and Possibilities

Jeffrey A. Rydberg-Cox <rydbergcoxj_at_umkc_dot_edu>, University of Missouri-Kansas City

Abstract

Incunabula, or books printed before 1500, are extremely difficult and expensive to convert to digital form. The primary challenges arise from the use of non-standard typographical glyphs based on medieval handwriting to abbreviate words. Further difficulties are also posed by the practice of inconsistently marking word breaks at the end of lines and reducing or even eliminating spacing between some words. As such, these documents form a distinct genre of electronic document that poses unique challenges for conversion to digital form. From 2005–2007, the Preservation and Access Research and Development Program at the National Endowment for the Humanities funded a study to explore methods for digitizing these difficult texts. This paper describes some of the results of that project. [1]

Introduction

Incunabula and other early printed works pose intriguing challenges and test cases for large-scale mass digitization projects such as Google Books[1] and the Open Content Alliance.[2] While there are many different challenges related to the digitization of early printed books such as encoding layout information, marginalia, hand-added marks to denote the structure of works, etc., the project described here focused on only one class of problem. Our primary aim was to study the challenges associated with representing in digital form the complex and non-standard typefaces used in these texts to abbreviate words in imitation of of medieval handwriting practice. Although there are many different types of abbreviation, the most prevalent is the use of a macron over vowels to indicate the removal of the subsequent nasal consonant[3]. These glyphs can represent different sets of characters in different contexts. For example, the string “ê” can be used to designate “en” or “em”, while a stylized variations on the letter “q” can denote many different common Latin words such as “que” and “quod”, and stylized variations of the letter “p” to denote the common prefixes “per” or “pro”. There are also specialized characters for some very common words. 1

Figure 1. Sample Brevigraphs

In addition to non-standard glyphs, two other early modern typographical practices make incunabula difficult to digitize. First, word divisions or spaces between words are often unclear; it can be difficult for data entry contractors to tell where one word begins and the next word ends. 2

3

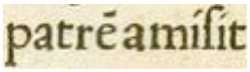

Figure 2. Sample Unclear Word Division

Second, words often break across line endings without hyphenation or any other mark. For example, the first three characters of a word can appear on one line with the final three characters on the following line and no typographical indicator of the word break. This practice can vary from line to line within a text; one line may be hyphenated while the next line will have a word break that is not hyphenated. 4

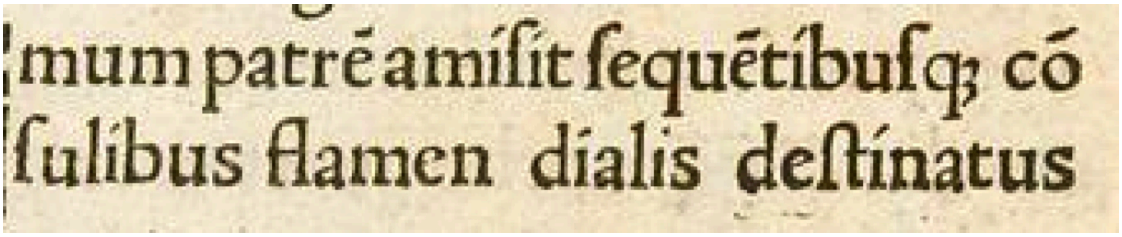

Figure 3. Sample word break across line ending

These examples of the most common abbreviations only begin to describe an extremely complex typographic system. For example, the 1488 edition of the *Stultifera Navis* was printed using some 168 unique glyphs while Pincius' 1497 Latin translation of Diogenes' *Vitae et Sententiae Philosophorum* contains some 156 unique glyphs. With the additional complexities introduced by varying practices from press to press across Europe and even among compositors at a single press, the number of possible variations expands exponentially. 5

These features of early typography occur at varying rates in different texts, but they appear so frequently in every early printed text that any digitization project must try to account for their presence. For example, the Archimedes Digital Library Project to digitize important early printed source texts for studying the history of mechanics created a library of sixty texts printed between the years 1495 and 1691.[4] In this broad corpus of texts, on average there are three to five abbreviations on every page. As one moves earlier in the history of print, abbreviations become much more common and correspondingly more difficult to deal with. For example, the Stultifera Navis (1488) has 313 pages and 12,404 abbreviations, an average of 39.8 abbreviations per page; the *Legenda Aurea* (before 1478) has 829 pages and 120,261 abbreviations, an average of 145 abbreviations per page; and the *Vitae et Sententiae Philosophorum* (1497) has 190 pages and 52,267 abbreviations, an average of 276 abbreviations per page. 6

Because of the prevalence of these glyphs, incunabula cannot be processed using OCR software. Commercial OCR programs produce almost no recognizable character strings, let alone searchable text. Sravana Reddy and Greg Crane have shown that ABBY FineReader can recognize only 84% of the glyphs on a page and that the open source GAMERA system can be trained to recognize only 80% of the glyphs in texts from our testbed [Reddy 2006]. Other methods must be explored. 7

Methods

When considering methods for digitizing early printed books, it is first essential to ask how much functionality end-users require from a digital facsimile of an incunable and study how much human intervention is required to create a text that end-users might find usable. We evaluated the costs and added value associated with different approaches to digitizing these documents. These approaches can be roughly organized into five groups: 8

- **Image Books:** Image books are simply page images that are scanned or photographed and placed online. A user can browse the pages sequentially, but there is no text to be searched — only images to be viewed.
- **Image Books With Minimal Structural Data:** This structural data includes section headers, chapter

headings, indexes, et cetera. These allow for some navigation of the text, but still not full text searching.

- **Image Front Transcriptions:** Image front transcriptions were pioneered by projects such as Making of America at the University of Michigan and Cornell. In this format, page images are run through an optical character recognition process and then the uncorrected OCR is used for searching, but the end user is only ever presented with the page image. This means that some of the benefits of searchability are made available to the end user while not incurring the expense of actually going through and cleaning up the OCR text to a point where it would be presentable to a human user.
- **Carefully Edited and Tagged Transcriptions:** These sorts of texts are the kind that you might find in a digital library such as Bartleby or the Perseus Project. These are texts that exist in the public domain that have been typed or processed via OCR. A scholar carefully checks the text against the original to make sure it is accurate and places it online. Generally, these texts are marked up in XML (projects that began before the creation of XML might also have texts in SGML). These texts are also marked up according to a DTD such as DocBook or the Text Encoding Initiative.
- **Scholarly and Critical Editions:** These digital texts replicate the work that scholars have done with printed text for hundreds of years. Scholars will choose an author, examine all his or her works, study all the variants and create an authoritative version of a text. They will replicate in digital form all of the standard footnotes and critical apparatus that one would expect in a printed scholarly edition of a text.

9 In our project, we initially planned to create tagged transcriptions of many different sample texts. As we considered these different genres, we decided that it would be much more interesting to create sample texts in each of these genres except for a full scholarly critical edition. This would allow us to see and evaluate first hand the costs and benefits of the different approaches. We created image books of Petrarch's 1492 De Remediis Utriusque Fortunae (Cremona, Bernardinus de Misintis and Caesar Parmensis; text images provided courtesy of National Library of Medicine) and Isidore of Seville's 1472 Etymologiae Liber IX (Ausberg, Gunther Zainer; text images provided courtesy of National Library of Medicine). We created image books with minimal structural data of Pliny the Elder's 1472 Naturalis Historiae Liber (Venice, Nicolas Jenson; text images provided courtesy of Linda Hall Library). We created image front transcriptions of Sebastian Brant's 1488 Ship of Fools (Lyons, Jaques Sacon; text images provided courtesy of Harvard University Countway Library) and Jacobus de Voragine's pre-1478 Legenda Aurea (Ulm, Johann Zainer; text images provided courtesy of Conception Abbey Seminary Library). We carefully transcribed and edited Suetonius' 1494 Vitae XII Caesarum (Milan, Leonard Pachel text images provided courtesy of Conception Abbey Seminary Library) and Diogenes Laertius' 1497 Latin translation of the Lives of Eminent Philosophers (Philippus Pincius for Benedictus Fontana; text images provided courtesy of the Universiyt of Missouri Kansas City Libraries.)

10 As one moves from image books to critical editions, the cost of the digital facsimile predictably increases because each element of human intervention introduces new costs. Capturing an image is the least expensive minimum baseline that all other editions build on; structural metadata adds human labor and expense beyond the cataloging of each image. Assuming that a text is amenable to OCR, image front transcriptions are only slightly more expensive and might even be cheaper than image books with manually created minimal structural metadata. The creation of a scholarly or critical edition lies at the other end of the cost spectrum; this is the sort of work that might take many years or even represent a scholar's lifetime achievement. The biggest jump in cost comes in the move from an image front transcription to a carefully edited and tagged transcription. Although preparing a page for OCR adds additional labor, manual data entry costs substantially more; when an editorial team must also be paid, costs escalate rapidly.

11 It is least expensive, of course, to produce digital images of a book and not provide any searchable text; the ability to search a text, however, is such an important function for most users that the extra labor and expense is justified. An optimial compromise between cost and functionality seems to emerge in the image-front model pioneered by projects such as the Making of America and now being implemented on a broad scale by Google Book Search and the Open Content Alliance. This approach is successful because it allows for the cost-effective digitization of large numbers of books while also providing the functionality and readabilty that most users expect. Indeed, most people who use digital texts expect to be able to search them, but they are not necessarily displeased to find a page image as opposed to a typed text transcription. While the findings of Reddy and Crane show that incunabula cannot yet be processed using this

sort of workflow, these large scale projects are extremely intriguing. Ultimately, if the problems of typography and layout can be resoloved to the point that these texts can be processed automatically, it would be possible to imagine the creation of similarly large corpora of early printed books.

12 The image-front approach to digitizing modern books should be applicable to early printed books because many of the high cost components are the same regardless of the source text. The Open Content Alliance claims that their method has moved costs from twenty dollars per page to ten cents per page. In our project, we had similar costs; when creating a human-edited searchable transcriptions, we spent between twenty and thirty-five dollars per page. These costs broke down as follows. We spent between two and ten dollars for each photograph of a page. The cost here depended primarily on the physical location of the book. Many of the books that we photographed in this project were physically located at libraries in Washington, D.C. and Boston, so we had to factor in our travel expenses for those images. We were able to photograph the books located in Kansas City for roughly two dollars per page. We then paid two dollars to six dollars per page to create character sets and to have the text typed. We paid one dollar and twenty-five cents per megabyte and used character sets to identify the unusual and non-standard characters that the data entry contractors encountered. In general, we found that it was possible for a human to edit the transcriptions and expand the abbreviations at a rate of two pages per hour. In our work, each page was checked twice resulting in a cost of ten to twelve dollars per page for tagging and proofreading. To this must be added supervision, training and project administrative costs, and institutional overhead rates. In order to reduce these costs and enable large scale digitization projects for early printed works, the primary aim must be to reduce tagging and proofreading costs. Because of the fragile nature of early printed books, manual photographers cannot be replaced by robotic scanners, but even with this expense our work suggests that it should be possible to create an image-front edition for somewhere between five and six dollars per page.

Data Entry Methodology

13 Once the decision has been made to digitize the text rather than displaying page images or page images with minimal metadata, the next issue to be faced is how to address the numerous characters and glyphs that cannot be represented in current computer encoding systems including Unicode. It was necessary, therefore, for us to develop a method that data entry contractors could use to represent these characters as they were typing our texts. The data entry process begins with a manually created catalog of every brevigraph that appears in each printed book (illustrated in Figure 4).

Figure 4. Sample Catalog of Brevigraphs

14 In this process, a unique entity identifier is assigned to each non-standard character that data entry personnel use to represent that glyph in a text. For example, the opening lines of the 1494 edition of Suetonius' Vitae XII Caesarum is printed as follows:

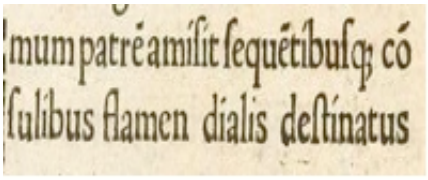

Figure 5. Sample Lines from Suetonius. *Vitae XII Caesarum.* Milan: Leonard Pachel. 1494

Using the catalog of brevigraphs, this line would have been entered by data entry contractors as follows: 15

```
patr&e1; amisit sequ&e2;tibus&q1; co&c1;sulibus flamen dialis destinatus
```
16

If the goal of this process is to create a carefully tagged TEI-conformant text that reflects the typography of a printed page, this product of raw data entry could be tagged as followed using the TEI abbr tag from the TEI P4 guidelines to represent each different glyph[5]. 17

```
patre<abbr type='e1'>m</abbr> amisit; seque<abbr type='e2'>n</abbr>ntibus<abbr type='q1'>que</abbr> co<abbr type='c1'>n</abbr>sulibus flamen dialis destinatus
```
18

Possibilities

Because the expansion of these abbreviations is an extremely time-consuming and painstaking task, we developed three tools to facilitate the tagging process. These tools suggest possible expansions for Latin abbreviations and brevigraphs, help identify words that are divided across lines, and separate words that are joined as the results of irregular spacing. All three programs can return results in HTML for human readability or by XML in response to remote procedure call as part of a program to automatically expand abbreviations in these texts. 19

The abbreviation expansion tool uses regular expressions to search all attested Latin forms in the Packard Humanities Institute Database of Latin, the Perseus Digital Library and our collection of texts. The program can return results in HTML for human readability or by XML in response to remote procedure call so that texts can be automatically tagged. 20

Figure 6. Abbreviation Expansion Tool

The unmarked word break tool takes broken words and searches that same lexical database for possible word combinations. This tool also provides its results by HTML and XML. 21

Figure 7. Unmarked Word Break Tool

Finally, the unclear word division tool takes a string and breaks it into pieces and searches for combinations that match attested forms in our lexical database and once again, the interface is provided for human readability in HTML and for use by a remote procedure call via XML. 22

Figure 8. Unclear Word Division Tool

The results are sorted by frequency to provide the editor with a sense of the most likely result. Although we conducted some experiments to see if n-grams or Hidden Markov Models can provide better weightings than raw frequencies, they did not provide more usable results for human editors. If we were to move to fully automatic expansion of these abbreviations, these weighting techinques would come into play. 23

Conclusion

Our work on this project suggests that it is possible to begin to conceive of a large-scale project to create image front editions of early printed texts with uncorrected manual data entry available for searching. The biggest expense in our workflow lies in the cost of having a human editor tag the abbreviations and a second editor proofread that work. Clearly, these costs can be reduced by creating image front editions on the model of the Open Content Alliance or the Google Book project. It would be necessary for us to replace the uncorrected OCR in their workflow with the uncorrected manually typed text from our workflow. Users could then search this uncorrected text and actually read from the page image. Because of the high rate of abbreviation, there is a price to be paid in terms of search precision. However, the precision/recall trade-off would not be such that it would render the text unusable. In the texts we studied for the NEH project, some 45-50% of words have no brevigraphs or abbreviations and another 25% of the abbreviations can be unambiguously expanded into a single word. The decrease in precision is introduced in the remaining 25%; of these, 10% have two or three expansions, 7.5% have more than three expansions, and 7.5% cannot be resolved. These numbers suggest that there would be a substantial increase in precision over recall for only a small percentage of searches. These results suggest that it would be possible to use these tools to create image front transcriptions. We could use our tools to automatically correct our transcriptions and perhaps have a single human editor check the corrections at a rate of four to five pages per hour. End-users would be able to search this transcription and read the page images. 24

While this approach advances our ability to undertake a large scale project to digitize early printed books, there is still some missing functionality. One of the advantages of a transcribed text is the ability to create automatic hypertexts. For 25

many years, the Perseus Digital Library has used its morphological analysis engine to create texts that are automatically annotated with morphological and syntactic data. In the Perseus interface, a user can click on a word to discover its part of speech and the lexical forms from which it could be derived. The interface also provides links to dictionaries and grammars that facilitate further inquiry.[6]

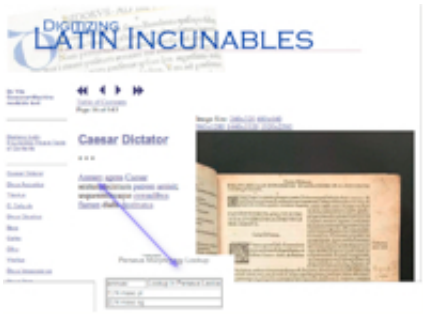

Figure 9. Sample Interface Showing Page Image and Transcription with Morphological Links

26 These tools allow student and scholars to read texts in the original Greek or Latin even if they do not have expert knowledge of these languages. These tools would be particularly useful for early printed books because these texts are essential for the study of almost every aspect of early modern culture. Many of them, however, have not been translated and many scholars who study this period are not afforded expert training in all of the languages that they may need to utilize to study this period.

27 The addition of this functionality might justify the substantial additional cost of creating a transcribed text. It also creates a compelling question for our research. If our image front transcription could also contain pixel coordinates for individual words, it would be possible to access the automatic hypertext via the page image. Users could click on the page image to see the raw text, the automatic expansions suggested by our tools and also the morphological and syntactic information from the Perseus morphological analyzer. This model is slightly more expensive than one where texts are just searched, but the image is not clickable. The data entry costs must increase because the typist will capture the pixel coordinates for each word. At the same time, however, editorial costs will fall because the transcription will only be used for searching with the word coordinates used for morphology and, in fact, we would hope that the text could come back from our data entry contractor and be used almost as is with minimal human intervention.

28 How do we analyze the costs and benefits of this approach? The idea of “just in time conversion” provides us with the best model for analysis. John Price Wilkins, in a 1997 article entitled “Just in Time Conversion, Just in Case Collections,” argues that digital library usage patterns mirror those of traditional libraries.[7] In this usage pattern, many materials can sit unused for very long periods of time punctuated by a period of high and intensive use. This pattern holds in the digital collections we have created for other projects. In some semesters, many texts will not be used at all, but one or two will be used very intensively. To the extent that this pattern holds larger collections of lightly edited text seem to provide more benefits to a broader audience than smaller collections of closely edited transcriptions or critical editions.

29 Further, using this model does not preclude the creation of a more carefully edited transcription or critical edition. Rather, it can provide the foundation for these texts. In this model, the raw or lightly corrected text can be provided to scholars with the expertise and inclination to create a more carefully edited version. This text could be released under a Creative Commons license that requires that the improved text to be returned to the scholarly community. For example, we could initially publish an uncorrected image front edition of the Legenda Aurea. A scholar working with this text could download the uncorrected transcription and devote their expertise to creating a better electronic transcription of the text. Under the Creative Commons license, they could use this as they wished and they would also return it to us and, ultimately, improve the entire digital library.

Notes

[1] The work described in this paper was completed by the “Approaching the Problems of Digitizing Latin Incunables” project funded by the National Endowment for the Humanities Division of Preservation and Access. The material in this paper is drawn from the project application, internal technical reports, grant project reports and the project descriptions included in [Rydberg-Cox 2003] and [Rydberg-Cox 2005]. Much of this work was inspired by Ross Scaife and his work building a corpora of Latin Colloquia. I am deeply grateful for Ross's comments, advice and support. A version of this paper will also be published as part of the project web site.

[1] [Google Books]

[2] [Open Content]; [BBC 2005]

[3] This discussion appeared in the original proposal, and was summarized in [Rydberg-Cox 2005]

[4] [Archimedes]; [Archimedes Templates]

[5] When this project was planned in 2004, we based our work on the TEI P4 guidelines and elected to use the `abbr` tag in our work. The new TEI P5 standard released in November of 2007 now has much more complete guidelines for working with character sets that are not represented in the Unicode standard. Although these guidelines appeared too late to inform the work of this project, the guidelines would need to be incorporated into any further work on these documents.

[6] http://www.perseus.tufts.edu; [Crane 1998]; [Crane 1991].

[7] [Price-Wilkin 1997]

Works Cited

Archimedes *Archimedes Project*. Harvard University. http://archimedes.fas.harvard.edu/.

Archimedes Templates *The Archimedes Project Digital Research Library*. http://zope.mpiwg-berlin.mpg.de/archimedes/archimedes_templates.

BBC 2005 BBC. *Microsoft scans British Library*. BBC, November 4 2005. http://news.bbc.co.uk/2/hi/technology/4402442.stm.

Crane 1991 Crane, Gregory. “Generating and Parsing Classical Greek”. *Literary and Linguistic Computing* 6: 4 (1991), pp. 243-245.

Crane 1998 Crane, Gregory. “New Technologies for Readings: The Lexicon and the Digital Library”. *Classical World* (1998), pp. 471-501.

Google Books Google. *Google Books*. http://books.google.com/.

Open Content Open Content Alliance. *Open Content Alliance*. http://www.opencontentalliance.org.

Price-Wilkin 1997 Price-Wilkin, John. “Just-in-time Conversion, Just-in-case Collections: Effectively leveraging rich document formats for the WWW”. *D-Lib Magazine* (1997). http://www.dlib.org/dlib/may97/michigan/05pricewilkin.html.

Reddy 2006 Reddy, Sravana, and Gregory Crane. “A Document Recognition System for Early Modern Latin”. Presented at *DHCS 2006*. *Chicago Colloquium on Digital Humanities and Computer Science: What Do You Do With a Million Books?* (2006).

Rydberg-Cox 2003 Rydberg-Cox, Jeffrey A. “Automatic Disambiguation of Latin Abbreviations in Early Modern Texts for Humanities Digital Libraries”. Presented at *JCDL 2003*. *Proceedings of the 2003 Joint Conference on Digital Libraries* (2003), pp. 372-272.

Rydberg-Cox 2005 Rydberg-Cox, Jeffrey A. *Digital Libraries and the Challenges of Digital Humanities*. Chandos Press, 2005.

Seattle 2005 “Alliance Aims to Digitize Classic Books”. *Seattle Times Books Section,* (Monday, October 24, 2005.).