A Mass Digitization Primer

Juliet Sutherland

Abstract

Many people are talking these days about "digitizing books." But what does that really mean? This paper describes different kinds of digitizing, the pros and cons of each, and suggests a layered structure for understanding "digitization."

The Digital Age has brought many challenges for librarians. Most obvious are all the issues concerning "born-digital" material that academics and the general public are generating. However, born-digital material is not the focus of this paper. A parallel effort to "digitize" the information currently locked up in print on physical paper has also emerged. There has been a lot of press recently about book digitization projects, as well as numerous predictions about the amazing things that can be accomplished when most of the world's older print materials are available online. In order to evaluate these claims, we must first understand what is involved in achieving them. And a large part of that is understanding the many meanings of digitizing books.

Page Pictures

The simplest meaning of digitizing books is that digital pictures have been taken of the book pages. Page images are stored in one of several common graphical file formats and someone who wants to access the book will look through these page pictures. This is very similar to storing a book on microfilm or microfiche, except that the digital form does allow for some additional functionality. Page images can be captured using either a scanner or a digital camera. For the purposes of this discussion, the method makes no difference, since the end result is the same.

LIBRARY TRENDS, Vol. 57, No. 1, Summer 2008 ("Digital Books and the Impact on Libraries," edited by Peter Brantley), pp. 17–23

Even saying that a collection has page images really doesn't provide much information. Some page-image archives provide exceedingly high quality, high resolution, color images. Others provide only black and white reproductions of questionable quality. Some page-image archives are meant to be archives for rare and fragile material, while others are not intended to provided archival-quality images.

What Page Pictures Are Good For

At their very best, page images can provide an experience that is extremely close to the physical reality of the book. See for example, the Posner Memorial Collection at http://posner.library.cmu.edu/Posner/, which uses very high quality, carefully produced page images. Page images show the exact layout on the page of text and illustrations, can capture the full colors and relative sizes of illustrations, show the fonts used for the text, and can even give a sense of the physical characteristics of the paper. A page image will show handwritten notes, foxing (age spots), tears, stains, uneven printing, and other artifacts that are unique to the physical item that was copied. Very high-quality images can also be used to print physical facsimiles of the item.

Having the high-quality image available in digital form on a network allows access to anyone who is also on the network. No need to spend the time involved in accessing a physical copy. Also, no additional wear and tear on the book.

In some cases, where an image of the "text" is as or more important than what the "text" says, an image archive may be the ideal and final solution. See, for example, the Cuneiform Digital Library Initiative (www.cdli.ucla.edu), Early Manuscripts at Oxford (http://image.ox.ac.uk), or any of the many other digital archives of rare and fragile written material.

Shortcomings of Page Pictures

At their worst, page images may not reproduce illustrations, may be very poor copies of the original, and may be incomplete. Consistently getting high-quality page images is more difficult than it may seem. On the other hand, for many purposes, very close facsimile may be overkill and unnecessary. For converting the book to etext, as discussed below, stains, foxing, etc. are actually a hindrance rather than a help. A lower-quality page image that shows the text clearly but does not reproduce aging paper or other such details is actually preferable.

While page images are absolutely necessary for certain kinds of research, they have serious disadvantages for most other uses. High-quality page image files are large, taking a lot of space to store and requiring a lot of bandwidth to access. The researcher based in a developing country with problematic Internet connections may find them difficult or impossible to use. The commuter who wants to read something on a PDA while traveling will find them completely useless. And even the computer user who looks through them on a full-size screen may not be able to read the text without awkward zooming. While the continually dropping prices for data storage and bandwidth may eventually make the file size issue irrelevant, the difficulties of scaling the image so as to be readable on various size screens will remain. Most of the books being produced by the large imaging projects are of a size that can be comfortably read by most people on most computer screens. But shrinking the image to fit onto a smaller screen will quickly render the text unreadable unless the reader zooms in on the image, and zooming requires moving the visible area around on the page, which quickly becomes a nuisance. For books with bigger physical pages and/or with very small type, or for people whose vision is a bit impaired, even using a large monitor may not make the text legible.

Variation within Page Image Projects

Most of the large print digitizing efforts are focused on creating page images of the material that they scan. Google is one of the best known. Microsoft, the Open Content Alliance (OCA), Boston Libraries, and others are also scanning existing print material as quickly as they can. The Library of Congress already has an impressive amount of scanned material. In addition many libraries have made, or are making, specific collections available. Lists of most of these can be found at the Online Book Page (http://onlinebooks.library.upenn.edu/archives.html).

Examining two of the major efforts, Google Book Search and the OCA, shows quite clearly how volume trades off against quality. Google has scanned a huge number of books. But the volunteers at Distributed Proofreaders (DP) have consistently found that the quality of the Google scans leaves a great deal to be desired when completeness is important. Many books are missing pages or have page images that are illegible in various ways. Illustrations are usually poor quality, especially color illustrations that have been scanned in black and white. Thanks to the volume of books from multiple libraries that are working with Google, it is becoming possible to piece together complete books by using pages from different copies of the same book. In contrast, the scans from the OCA that are produced by the Internet Archive's SCRIBE technology are full color and are rarely missing pages. The tradeoff, however, is that the OCA has scanned far fewer books than Google.

Raw OCR

Google's purpose, however, is not to archive printed material, but rather to make it accessible via their expertise in search. Since pictures of pages cannot be easily searched, Google runs an optical character recognition (OCR) program on their page images. An OCR program is able to recognize pictures of letters and convert them into a form that computers recognize as letters. By storing the results of the OCR, along with information about where each word falls on a page, Google can answer a query by going to the exact page and highlighting the relevant word. Virtually all page image archives provide a similar function, using OCR in the background so that a query will take the inquirer to a book that contains the desired terms. Some will highlight the term on a page; others will simply show the page. When a page image archive says that "full search" is available, raw OCR is what is powering that search.

OCR technology has improved radically since it was first introduced. Commercial OCR of modem printed materials is now virtually perfect. However, the older printed materials that are being scanned by the major digitizing efforts present much greater challenges. Uneven inking, worn type, defacing of the text, unusual fonts, changes in printing conventions, and odd page layouts all combine to make OCR a challenge. As one example, most OCR programs use dictionaries to assist their results, but certain kinds of errors, which will not be caught by a standard spell-check, are common. At DP we call these "stealth scannos." For example, look carefully at the word "modern" earlier in this paragraph. The modem/modern confusion is common. Fortunately, since the word "modem" would be highly unlikely in texts from the early 1900s it can safely be automatically corrected for older texts. Far more complicated is misreading "he" for "be," and vice versa. The he/be misreads cannot currently be reliably corrected by an automated process.

Some other OCR challenges include punctuation, which is particularly problematic since the difference between a comma and a period is very small to begin with and is easily obscured by poor printing, and older printing conventions that often left a partial space between punctuation and its associated text. Since computers and OCR don't deal with partial spaces, these are often translated into full spaces. While many cases of spaced punctuation can be corrected automatically, some cases are ambiguous enough to require human attention. This is by no means a full listing of the common OCR errors encountered in older material but should suffice to make the point that high quality OCR output from old texts is difficult.

Corrected OCR

The next step in digitizing printed material is correcting the OCR so that the electronic form of the text matches what was printed on the page. Because so much of this cannot be mechanized for older texts, it is labor intensive and therefore very costly. None of the major digitizing efforts are attempting to take this step, nor can they afford to do so under current conditions.

Fully digital text allows a great deal of additional functionality. Digital text can be reflowed and rewrapped so that the text can be read on screens of any size. The character size can be increased for easier reading. Text-to-speech programs can use it for the visually impaired. Text can be easily copied for quoting or excerpt purposes. High-quality reprints become possible with digital typesetting in modern type. Search will work more accurately. Textual analysis as well as all kinds of interesting automatic linking becomes possible. Finally, much of the promise of online access to older books depends on semantic processing that in turn depends on having accurate text to work from.

Some scholarly programs have obtained funds for producing accurate etext for certain collections. The most common method used to make these texts is double-key entry, typically in a developing country where wages are cheap. In double-key entry, two people type in the text and the versions are compared to produce an accurate final version. This is based on the premise that two people are unlikely to make the same typos. Even with cheap labor, however, this is an expensive process.

One fascinating approach to this problem, called Recaptcha (http://recaptcha.net), has been developed by Luis von Ahn of Carnegie Mellon University. This process identifies individual words, particularly those about which the OCR is unsure, and presents the image of them to users as part of the human authentication process in accessing certain websites. The users type in the correct version of the word and if several people agree, that word can then be corrected in the text version. This is a very clever way to harness small snippets of time from lots of people to perform tasks that humans are good at and machines are not. For the moment, the Recaptcha project is working with texts from the Internet Archive and this process should greatly improve the quality of those texts.

Recaptcha does have some limitations. It can only ask humans about items that the OCR process has identified as words and about which the OCR engine is unsure. It is very common for OCR programs to overlook words that are either completely illegible or sometimes words that appear alone. It is somewhat less common for OCR to confidently recognize a "stealth scanno" but it does occur. Further, some individual words can be very hard to identify without surrounding context. Recaptcha does not address the potentially ambiguous issues of spaced punctuation. And finally, it doesn't address long-s, or other now unconventional forms of orthography. So far it only seems to have been used with English texts. Other languages will pose additional problems due to spelling reform and archaic usage.

Another approach to correcting raw OCR has been to make the text available to anyone for correction. In this scenario, anyone reading from page images is encouraged to make corrections to the accompanying text. This approach suffers from a variety of problems. First, it can be difficult to tell how much work has been done on any particular text. Second, there is no assurance that any text will be "finished," meaning that someone has at least looked at every page. Third, it is difficult to protect from vandalism, or, more insidiously, minor changes that can have large impact. And finally, even assuming good will, not everyone will make changes correctly and checking mechanisms are uncommon or unwieldy.

Distributed Proofreaders (www.pgdp.net) uses volunteer labor to produce extremely accurate digital transcriptions. Volunteers are shown an image of the physical page together with the current text associated with that page and are asked to correct the text to match the image. Each page is looked at three to five times by different volunteers. When the page level process has finished, a volunteer assembles all of the pages and does more checks that are most easily done across the entire text. While the result of all this effort is very accurate, it is not a speedy process. The primary Distributed Proofreaders site is currently producing about 220 texts per month. While that number grows as the volunteer base grows, it will never match the tens of thousands of texts per month that are being scanned by the major digitizing efforts such as Google and the OCA.

Semantic Coding

The final step in making digitized texts most useful is encoding the semantics contained within them. This can be as simple as identifying chapter titles or as complex as identifying whether a particular instance of the word "Washington" refers to the person, the city, or the state. Many of the more futuristic predictions for what will come from digitizing that portion of humanity's knowledge, which are currently locked up on paper, require this sort of semantic identification. It is not yet clear how semantics will be identified, expressed, and used. What is clear, however, is that information science researchers are developing ever more ingenious ways of mining data for this kind of information.

An Example: The Biodiversity Heritage Library (BHL)

Having discussed the various activities that might be included in digitizing texts, applying them to an example should be instructive. Disclaimer: the author has no association with this project other than as an interested observer.

The Biodiversity Heritage Library (www.biodiversitylibrary.org) says about themselves:

> Ten major natural history museum libraries, botanical libraries, and research institutions have joined to form the Biodiversity Heritage Library Project. The group is developing a strategy and operational plan to digitize the published literature of biodiversity held in their respective collections. This literature will be available through a global biodiversity commons.

What makes this project particularly interesting is that unlike most sciences, where older research papers, books, and periodicals are only of historical interest, species descriptions from 100 or more years ago may well remain authoritative, and in some cases may be the only extant descriptions. Making this material available will therefore directly benefit current research, both by making the material easier to find, and by making it available to researchers in developing countries, or, indeed, to anyone who does not have the budget to visit the institutions that hold the originals of these papers.

The first decision this project had to make was how best to get page images from the original material. They chose to use the Scribe technology from the Internet Archive, which is also storing all the page images for them. This means that they will have archival quality, color images, with raw OCR behind them. These images alone will be a huge improvement in making these materials widely available.

Raw OCR, however, might very well misread species names, potentially very confusing if the names differ by only one or two letters. Text in all capitals, as might occur in the title of a paper or chapter, and text in italics, which is often used to indicate a species name, are particularly prone to mis-reads. With these kinds of errors researchers may miss some relevant material, or perhaps get far too many false positives when doing a search. In an ideal world, BHL would have corrected text available as well as the page images.

Taking matters one step further, if area specialists were able to include semantic tags in the texts, marking those items that might be of interest to researchers, the entire text base would become a giant database on which complicated and interesting queries might be made.

As matters are right now, corrected text, and semantic tagging within the text, are too labor intensive to be economically feasible, although technological progress will undoubtedly help to address these problems over time.

Summary

Digitizing information should be thought of as a multilayer, multistep process. It starts with images of pages, progresses through raw OCR to corrected text, and finishes with encoding of semantic information contained within the text. Each step provides more utility but at additional cost.

Juliet Sutherland is the chief PTB (Powers That Be) at Distributed Proofreaders (www.pgdp.net), a completely volunteer website that transcribes public domain texts into electronic form. In addition to her administrative duties, she has been the project manager for over 1,200 of the more than 12,000 titles completed there.