Working with Printed Text and Manuscripts

Stephen Chapman
Harvard University Library

Look at the growing body of network-accessible books, journals, and archives from cultural institutions and commercial publishers and you will discover that electronic text is not all alike. Some collections are searchable, others are not; some have highquality color reproductions, others limit their content to black-and-white (1-bit) images; some support go-to-page and go-to-section navigation, and many simply provide page-forward, page-back functionality. Rather than present a single case study of one type of electronic text, this section presents a composite case study of the challenges raised by several types of text conversion and the guidelines that have emerged in response to them.

Since costs among all of these versions vary widely, the first job for the budgetconscious manager is to select the least-expensive electronic publication model appropriate to the collection(s) she or he has selected for digitization. Generally speaking, electronic text falls into three categories:

> ***Page Images*** These *digital photocopies* are created by scanning printed pages or microfilm. Page images are not searchable. They may be black-and-white, grayscale, or color. Assume that black-and-white (1-bit) page images are the least expensive products of text digitization, but be sure to account for the associated costs of the structural metadata that is needed in order to make the images suitable for browsing and on-line navigation.
>
> ***Full Text*** In order for printed text to become fully searchable electronic text (full text), the letters on the original pages must be translated to machine-processible ASCII. There are two ways to do this: either by typing from the original (known as *keying*) or by using an optical character recognition (OCR) program to convert page images to ASCII. The first process is manual, the second automated. Since keying can easily be ten times more expensive than scanning-plus-OCR, page images are often made to facilitate the creation of full text.
>
> When these two products (full text and page images) are made, there is the added advantage of being able to present a *facsimile* version of the original page -- with fonts, formatting, marginalia, and illustrations intact -- in response to a search. In other words, the ASCII is used to create an index for searching, and only the page images are delivered to the screen or printer.
>
> One might ask: If scanning and OCR are so much cheaper than keying, why consider keying at all? First, OCR is viable only for page images of machine-printed text. Handwritten originals must be keyed to become searchable. Second, OCR accuracy decreases as the complexity of originals increases (number of fonts, number of columns, illustrations accompanying text), and as the quality of the page images decreases. Therefore, if near 100% accuracy of searching is required, it might be less

expensive to key than to undertake the three-step process of scan, OCR, and correct OCR errors. Several reliable studies report that a trained technician can correct 6-10 pages per hour. Depending upon salary, this task alone could easily exceed the cost of keying.

Encoded Text, or Full Text with Mark-up This third publishing model for text conversion is the most expensive, but also the most functional and flexible in the online environment. Like plain full text, encoded text production requires keying or OCR of page images to create ASCII. The final step is to encode attributes of a given document by placing Standard Generalized Markup Language (SGML) tags around selected text. There are hundreds of SGML elements that can be used for encoding. The Text Encoding Initiative (TEI) Guidelines refer to a subset that has been used widely for publications in the humanities. Texts usually are encoded at one or both of the following levels: (1) structural: referring to divisions such as chapters within books, articles within journals, poems within anthologies; or (2) descriptive: referring to elements such as dates, names of persons or places, and occupations. When a properly configured search interface/application is coupled with an SGML database, encoding makes fielded searching possible (*e.g.,* find "slavery" in captions), and can also be used to control the presentation of the document -- including multiple representations if desired.

Note: It is not necessary to create page images in order to produce encoded text if (1) keying is an affordable approach to production, and (2) your goal is to present modern rather than facsimile pages to the screen or printer.

After deciding which electronic products satisfy the project requirements, the manager's second task should be to specify the outcome for the source materials after conversion. Since the printed originals are also products of text conversion, it is important to determine whether they should emerge exactly as they began or whether alterations are acceptable. It is significantly easier to create a project budget and plan of work if disposition decisions -- related to access policies, materials housing and location, and even deaccessioning -- are made at the outset.

Decisions about the appropriate outcomes for the source materials inform, if not determine, the handling guidelines for scanning. Materials that will be kept, particularly if they are to remain as circulating copies, may need to be assessed, cleaned, repaired, or rehoused at some point in the project. On the other hand, materials that will be moved to offsite storage or even discarded allow for a greater range of options in scanning techniques.

Questions about handling and disposition are particularly important for bound materials. Disbound pages, even when highly brittle, can either be scanned on flatbed scanners or can be automatically fed to sheetfeed scanners (with straight paper paths). In other words, it is much less expensive to scan pages than to scan books. As of 1999, production statistics gathered from a number of projects
indicate that although technicians can scan up to five pages per minute, they typically average between two and three. Auto-feed scanners, on the other hand, can scan two sides of a page in a single pass. Using the same output settings (*e.g.,* 600 dpi1-bit TIFF), these scanners produce 20 images per minute. Thus, assessments of source materials are critical because

whenever manual feeding (or page turning) is required, scanning prices are tied directly to labor costs. In this model, improvements in scanning technology can only result in better quality, but not higher speed. When auto-feeding is allowed, technology improvements can result not only in higher quality but also higher speeds, and therefore lower unit costs.

Decisions regarding appropriate handling are complex, and any method must be tested and confirmed with a sample of materials before undertaking full production. No single best system has emerged for text scanning. Auto-feed, flatbed, overhead, or even digital camera systems are all viable. When selecting a scanner or writing specifications for a service bureau, handling requirements should be specified first, then image quality and speed. Scanning software plays an important role in these areas. For example, the same input settings -- *e.g.,* 600 dpi 1-bit TIFF -- on different scanners will produce different results on output. Batch settings often distinguish high-price from low-price systems and are critical for high-volume applications.

Rules of Thumb

Although there are many variables associated with selecting the best methods to create page images and/or full text, there are fortunately some rules of thumb common to many text conversion projects.

- To minimize costs of creating and maintaining page images, 1-bit scanning with lossless compression should be used whenever possible; permitting the use of auto-feed scanners is the least expensive way to produce images of high enough quality for OCR, printing, and/or computer output microfilm (COM). Quality from all 1-bit scanners -- sheetfeed, flatbed, and overhead -- is the product of engineering (hardware, optics), software, and operator skill, so be sure to confirm that resolution requirements cited in one project work equally well for the materials and scanner you have selected in yours.
- When grayscale or color scanning is preferred for machine-printed text, use a scanner or digital camera with enough spatial resolution to capture the lines, edges, and other details of the source materials. Compare the costs and quality of line-array and area-array systems to determine which produces acceptable quality at the lowest cost. If OCR is required, fairly sophisticated image processing (following scanning) will be needed to generate 1-bit files from the grayscale or color scans.
- When conservation assessment and/or treatment is mandated for the source materials, conservators should participate in selecting the scanning equipment that will be used and in writing the handling guidelines for the project.
- Image quality and quality control requirements relate directly to the disposition of the source materials. Quality requirements will be higher for projects where reduced access to, or even replacement of, the originals is required. Costs, ironically, may be lower, since auto-feeding may be viewed as a more acceptable technique for these items than for unique materials in good condition.
- Costs of document preparation (excluding conservation treatment), metadata creation, and quality control are likely to exceed the cost of scanning, particularly for 1-bit imaging.

- Given the design of overhead scanners, as well as the limited depth of field in many digital cameras, bound volumes will be less expensive to scan if they can be opened fully (180 degrees). Text printed near or into the gutter margin is always difficult to capture -- as handling requirements increase, so will the costs.
- Oversize pages (particularly when the longest dimension is greater than 17") are always more expensive to scan. High-quality digital reproduction of text becomes more difficult with direct scanning; newspapers, for example, have routinely been microfilmed first in order to produce page images of adequate quality.
- Many image enhancement techniques, such as despeckling and deskewing, can be automated following scanning. Image processing is important not only to the appearance of page images, but also to their optimization for OCR.
- The structural metadata needed to organize page images may be created before, during, or after scanning. Given the idiosyncrasies of pagination and organization of many historic collections, one should expect these tasks to be manual, or semiautomated at best.
- Requirements for full text accuracy and depth of encoding result from a careful analysis of the source materials and consultation with the community(ies) interested in using the digital collections.
- The following table summarizes the decisions that have the most important impact on quality and cost in text conversion projects. Many guidelines have been proposed from case studies, and these have been generalized for the table. As discussed in other chapters, however, good management begins by setting goals, not by blindly following guidelines. Relate your decisions to your publication
- objectives and preferred outcomes for the source materials, and the scanning guidelines and costs will naturally follow.

KEY QUALITY AND COST DECISIONS FOR DIGITIZED TEXT		
Product	**Examples of Key Decisions**	**Guidelines**
Source Materials	*Handling* * Contact with glass permitted	All scanners are viable
	***** Bound volumes must be supported during scanning (opened less than180°)	Face-up scanning required, with appropriate cradle/book support
	Disposition * Maintain standard of access: return in original format to original location	Identify resources available for treatment. If staffing and funding are available, for example, to assess, disbind, and rebind materials, then compare costs of scanning pages versus scanning books before selecting best approach.

	***** Reduce access by changing circulation policy or by relocating	To save cost, auto-feed if feasible, but budget for necessary preparation material and rehousing costs.
	***** Severely reduce or even eliminate access by creating digital images of replacement quality and/or by disposing source materials after scanning	Requirements for quality control and metadata must be explicitly defined (consider use of technical targets); disbinding might be most appropriate in these circumstances.
	Preparation ***** Facilitate highest quality scanning at the lowest cost	Segregate materials into batches whenever feasible (*e.g.,* by size; or by content -- text, illustration, mixed, color)
Page Images	*Specifications for master (archival) images* ***** Achieve tone reproduction appropriate to source materials and output requirements	When black-and-white (1-bit) fails to capture essential information, use scanners that sample 12-bits per pixel and output at least 8-bits per pixel for grayscale and 24-bits per pixel for color.
	***** For machine-printed text, achieve detail reproduction needed to meet output requirements (screen, print, OCR for machine-printed text)	400-600 dpi commonly used; threshold and image processing capabilities also critical to image quality, especially for 1-bit images; post-scan enhancements can increase OCR accuracy
	*****For handwritten manuscripts and soft-edge type, such as photostats, achieve detail reproduction needed to meet output requirements (screen, print, zoom)	300 dpi minimum for 1-bit, 200-400 dpi minimum for grayscale and color
	*****Use open format	TIFF
	*****Use safe compression	Group 4 (lossless) compression for 1-bit, none for grayscale and color images
	***** Implement quality control program	Confirm that all files for object have been received, sequence is correct, metadata is complete and correct (100%); check image quality on screen, in print or both (sample)
	Specifications for delivery images (derivatives) * Print, computer-output microfilm (COM)	Master images (highresolution TIFFs), PDF, or Postscript
	***** On-screen images	Legibility generally achieved at 80-120 dpi; minimize file size by using fewer than the

		full 8-bits for GIF whenever possible (*e.g.,* 4-bit); if 8 to 24-bits are required, consider JPEG
	Specifications for navigation ***** Page-forward, page back	Include *sequence* field in image database, or embed sequence in filenames
	***** Go-to page	Include *page number* field in image database, or embed page number in filenames (the latter is generally a more expensive solution)
	***** Go-to section	Include *feature* or *feature code* field in image database, or mark- up full text (see below)
Full Text	*Specification for accuracy (characters only)* ***** 100%	Get prices for keying first, then conduct sample OCR test of page images
	***** less than 100%	Conduct sample OCR test of page images and review acceptability of output; avoid need to correct OCRgenerated text at all costs
Marked-up Text	*Specification for accuracy (characters and formatting)* ***** Fidelity to original required desirable	Keying/encoding may be the least expensive approach; test scanning/ OCR only if the original layout and fonts are relatively simple
	Specification for encoding * Accommodate attributes of materials in hand while using practices endorsed by broader community	Consult TEI LITE and create DTD to accommodate structural divisions and descriptive features in the texts in hand; local interpretations of the general guidelines are possible

Sources

Bicknese, Douglas A. *Measuring the Accuracy of the OCR in the Making of America.* Winter1998. http://moa.umdl.umich.edu/moaocr.html

Guthrie, Kevin M. "JSTOR: From Project to Independent Organization." *D-Lib Magazine* (July-August, 1997). http://www.dlib.org/dlib/july97/07guthrie.html

Morrison, Alan, Michael Popham and Karen Wikander. "Creating and Documenting Electronic Texts: A Guide to Good Practice." *AHDS Guides to Good Practice.* Arts and Humanities Data Service, 2000. http://ota.ahds.ac.uk/documents/creating/

Text Endocding Initiative (TEI). "The TEI Consortium Homepage." http://www.ctei-c.org/

University of Virginia Library Electronic Text Center. The Electronic Text Center: On-Line Helpsheets. http://etext.lib.virginia.edu/helpsheets/sgmlscan.html