LEVELS OF SERVICE FOR TEXT DIGITIZATION

Text reveals better than images the spectrums of *cost*, complexity, quality, and effort involved in digitization. The number of input and output formats to manage in this category is large; the number and variety of source items deserving of digitization is staggering.

Used here, *text* refers to any source item comprised of written pages. The support media for each page may be paper, parchment, vellum, photostat, as well as any photographic format. All types of writing or printing are included in this category, including handwriting. Text sources may be in any language; some may be printed in multiple languages.

Books and other text sources may be bound or unbound. Finally, they also may contain images and other nontext components (such as covers, endpapers). These multimedia sources also fall into the text digitization category.

Digitized text refers to three major genres of machine-readable data, each with spectrums of quality to achieve in production:

- Page images—digital images of each page (not searchable)
- Text or hidden text—plain text (ASCII), either keyed (transcribed) from each page, or generated from page images via optical character recognition (OCR) to yield alphanumeric data for indexing and searching, and sometimes for display
- Encoded text—text with descriptive markup (SGML, XML, HTML) to support multiple uses across applications (including navigation among different parts or features of multipage documents); transcribed or generated via OCR, frequently with correction, since encoded text is often displayed

In many library digitization workflows, both page images and either hidden or encoded text are produced as a cost-effective approach to preservation and access. Page images convey original layout and appearance; text facilitates keyword searching.

This chapter describes the program components necessary to produce discoverable, sustainable, and usable collections of digitized text for legacy collections of books, serials, manuscripts, archives, and other multipage works, as well as single-page source material—from small printed ephemera to oversize maps—whose meaningful content is primarily text, or text and line art.

Assuming that a library has already made appropriate program investments in digital library technology (Chapter 1) and digitization program management (Chapter 2), the baseline level of service for text digitization encompasses the staff, systems, and procedures necessary to manage all production tasks—from selection to delivery—for digitization projects.

Baseline text digitization services have the capacity to create page image or text products, with associated descriptive, structural, and administrative metadata that meet the following criteria:

- The digital reproduction is appropriately cataloged and discoverable, and the descriptive metadata are stored in a well-supported system

- The digital reproduction can be opened and rendered as a properly sequenced, navigable multipage object
- The digital reproduction is appropriately named to be identified by some type of inventory control mechanism (ranging for a printed list to a complex database)
- All files *corresponding to each page* are appropriately structured, stored, identified, and documented (with administrative metadata) for ongoing management
- The copy can be reliably delivered by the library's (or partner organization's) designated applications for networked access

Fulfillment of these minimum criteria—whether measured against local or community definitions of what is appropriate or good (see, for example, NISO's *Framework of Guidance for Building Good Digital Collections*)—are presumed to offer the potential for sustainability.

Levels of service above-the-baseline are required for any project that explicitly states requirements for pictorial quality in page images or that mandates production of text of high enough quality to be displayed.

The above-the-baseline services would also be instituted to support search and discovery with controlled vocabularies, and to increase the likelihood of sustainability through production of standards-compliant administrative metadata.

Baseline production services

Provided that downstream (post-digitization) systems are in place to store cataloging data and digitized text, and to make the catalog(s) and digital objects Internet accessible, key text digitization tasks (and their attendant standards, specifications, and best practices) requiring infrastructure are:

- Production of descriptive metadata (cataloging)
- Production of page images or encoded text (or both)
- Production of structural metadata

As with image digitization, baseline production infrastructure for text does not necessarily need to be extensive or costly to create an Internet-accessible collection of digitized text.

Provided that cataloging procedures are well-defined and implemented and that all digital products are adequately stored, librarians can—as several libraries have demonstrated—digitize newspapers (of small dimensions), sheet music, or pamphlets with a single flatbed scanner, and then deliver digital reproductions as portable document format (PDF) objects accessible via the Web.

Baseline production services for text digitization implicitly recognize the requirements to digitize *pages* and to digitize *works*. Particularly when production tasks are distributed among different specialists in assembly-line fashion, everyone engaged in the digitization project must understand and assimilate these two units of work.

When methodically planning the baseline text digitization infrastructure, the manager will account for tradeoffs in costs, quality, and sustainability in eight production arenas of text digitization, as presented below. The challenge is to configure systems, staffing, and procedures (services) that reflect and support program priorities over time.

Selection strategy

The point made in the previous chapter about digitizing images bears repeating: production managers and digitization technicians love homogeneity. When uniform, the following attributes of multipage text sources minimize the number of production set-ups and, consequently, minimize effort, project overhead, and costs:

- Size (dimensions)
- Format (print, 35mm black-and-white preservation microfilm, 105mm microfiche)
- Structure (bound, single sheet, one-sided, two-sided)
- Layout (within and among works)
- Language(s)
- Extent of nontextual components, such as illustrations
- Type and size of the most challenging meaningful details
- Range of meaningful tones or color (dynamic range)
- Condition
- Handling policy

Source materials with similar, if not uniform, attributes in all these areas can be handled in the same way and scanned to one specification. Setup time—per collection, per batch, and per page—is a meaningful cost component in digitization. Thus, selection strategies that can minimize differences among physical attributes of source materials without undermining project goals for usability promote highest-production, lowest-cost digitization.

By the same token, to simplify production and methods for sustainability, output digitized text objects would be of one product type: page images, encoded text, or page images + ASCII or encoded text.

Within this product type, each work would be comprised of the same component parts. Page images would be of a single format, compression type, and color space, and they would have the same number and type of related delivery images or ASCII pages for indexing and searching.

Encoded text, when present, would be of one format created to one set of rules. The type and extent of structural and administrative metadata also would conform to a single profile.

Unfortunately (from the perspective of digitization), text collections are heterogeneous, particularly when the collection spans a date range greater than 10 or 20 years, and especially when different source types (books and serials) are selected. Differences in original print quality and effects of aging also are common.

Ease of digitization should not be ranked ahead of intrinsic values (artistic, documentary, evidential, intellectual, etc.) when selecting text to digitize.

Stakeholders in any project should be aware, however, of the tradeoffs between homogeneity, which promotes high production and relatively low cost of digitization and sustainability, and heterogeneity, which increases costs by increasing numbers of production batches and preservation profiles.

Product choice and specifications development

Too many viable ways to digitize text exist to infer specifications from cursory or even careful examinations of source material. A typical book printed after 1890, for example, likely has a color binding (with or without an illustration), one or more screened (halftone) or photogravure illustrations, and a variety of fonts and layouts. It might also have a table of contents or index, and will likely be subdivided into multipage sections (such as chapters).

Many titles now worth digitizing were previously worth microfilming. Thus, candidates for digitization might exist in two formats: original print and surrogate 35mm preservation microfilm, begging questions of selection policies or preferences when multiple formats are identified.

Digitization specifications for books are product-driven. Unlike the Model T approach to choosing film stock for reformatting—"You can have any color you want as long as it's black[-and-white 35mm microfilm]"—digitization technologies present several viable options for books and other multipage sources:

- Librarians can digitize volumes to generate black-and-white page images for text and illustrations, and accompany these page images with hidden OCR-generated text for searching, as the Cornell University Library has done for over 1,500 volumes of Core Historical Literature of Agriculture.
- Althernatively, they can produce a highly-accurate SGML or XML encoded text transcription, presented to the user in HTML or e-book formats (or both), as the University of Virginia Electronic Text Center has done for 70,000 titles since 1992.
- Or if the works were deemed to be of high intrinsic and artifactual value—such as first editions from major American authors—librarians could digitize all pages in color and produce highly accurate encoded text transcriptions, offering the user a choice of format to work with, as the Massachusetts Historical Society did for its *Adams Family Papers Electronic Archive*; and the University of Virginia Library has done in partnership with ProQuest/Chadwyck-Healey for 886 volumes and 199 manuscript titles in their *Early American Fiction* project.

Core Historical Literature of Agriculture, http://chla.library.cornell.edu

University of Virginia Electronic Text Center, http://etext.lib.virginia.edu

Massachusetts Historical Society, www.masshist.org/adams

University of Virginia Library, http://etext.lib.virginia.edu/eaf

As a core component of baseline service, an operational text digitization program has processes in place to determine, at a high level, which product types it will invest in and support. Two factors greatly influence these decisions: user needs, and downstream infrastructure for storage, management, and, especially, delivery.

As with image digitization, starting at the end to develop project specifications is highly recommended. And, again, outsourcing digitization increases costs of developing specifications, particularly for text digitization where vendors must have explicit requirements for digital imaging, text, text encoding, and structural metadata. (See, "Specifications Development" in Chapter 3 for more details.)

Metadata production

As described in the previous chapter, descriptive metadata help to sustain digitized text collections by developing the support base of users who might partner in preservation activities—from advocacy to contributing tangible resources.

When libraries digitize published monographs and serials, they now have an opportunity to extend this user base to every library that owns a copy of the work by *registering* the digital reproductions in the OCLC "Registry of Digital Masters."

In 2001, the Digital Library Federation articulated both a general case and functional requirements for such a registry, underscoring the needs to prevent inadvertent duplications of effort (for example, digitization of a work more than once), to advertise the availability of a persistent good copy, and potentially "to seek economies by sharing responsibility for maintaining digital resources for the long term."

DLF/OCLC Registry of Digital Masters,
www.diglib.org/collections/reg/reg.htm

Libraries seriously committed to sustaining digitized texts would incorporate the "Registry of Digital Masters Records Creation Guidelines" into cataloging workflows. This additional cost to create registry metadata would be recovered by time saved to locate existing digitized versions of items being considered for digitization from a librarian's own library.

Although standards and best practices for descriptive metadata are beyond the scope of this report, note that creating an electronic version of a finding aid (from paper) for archival materials and special collections constitutes a separate text conversion project within a text digitization project.

Given the strong interest in digitizing primary source material, a library would likely have to, in time, integrate finding aid creation and conversion capabilities into its baseline services for text digitization.

Both finding aids and bibliographic records need to be carefully structured (potentially revised) when only parts of a collection or series (such as a multiyear, multivolume serial) are being digitized. It is important to verify that the metadata records have been created to sufficient depth (granularity) to link logically to the parts of the whole that have been digitized.

Structural metadata

Books are remarkable products of collaboration, reflecting the discrete specialties of design, papermaking, printing, graphics, gathering and sewing, and binding. Digitized texts are the modern equivalents of their printed progenitors, bringing together the individual products of digital photography and scanning, sequencing and paginating, text production and encoding (sometimes through scholarly analysis), and, finally, interface design.

Structural metadata assemble pages (whether page images or individual ASCII files) into functional digital objects. Computers are literal. Neither page images nor simple (unencoded) text formats possess inherent attributes that command an application to turn pages, go to a specified page number, or jump to a specific section—key behaviors, in the parlance of the University of California at Berkeley's *Making of America II* team, used to champion electronic editions over microform or photocopy surrogates.

Key issues for the operations manager to confront are standards, tools, and costs. Structural metadata, like any other type of mark up, can be produced to various levels of granularity, so costs are driven in part by specifications for extent of metadata.

Structural metadata can be gathered before digitization from the source item in hand, or afterward from page images. In the former case, metadata are always manually created by typing attributes such as page numbers or chapter titles into a database, spreadsheet, or HTML form.

When produced after scanning, structural metadata may either be created manually or automatically. In the case of high-end applications optimized for certain types of text (such as newspapers), layout recognition software with OCR and mark up features can automatically locate and record structural metadata to varying degrees of accuracy.

Selected examples of metadata production tools—most configured to generate metadata either in simple tab-delimited format, or compliant with standards such as Text Encoding Initiative (TEI) SGML, Making of America II (MoA II) XML, and, most recently, Metadata Encoding Transmission Standard (METS) XML—are listed below:

Commercial products

- Agfa BSCAN Capture Software, www.agfa.com/mds/document/capturesys/scansoftware/bscan
- CCS DocWorks METAe (high-end), www.ccs-gmbh.de/index_e.html, http://meta-e.aib.uni-linz.ac.at/index.html
- DiMeMa, Inc.'s CONTENTdm, http://contentdm.com
- informatik inc.'s Infothek Docudex software, www.informatik.com/docuthek.html
- Olive Historical newspaper collection software (high-end), www.oclc.org/olive/default.htm

Examples of locally developed tools

- California Cultures Project (Web-based form), http://calcultures.cdlib.org/project_manual/chapter1.html#anchorwebform
- Cornell Institute for Digital Collections, ImageTag for document structuring and tagging, http://cidc.library.cornell.edu/source

Implementations of customized solutions and adaptations of commercial products to fit local needs both depend on programming resources within the digitization operation. At this writing, structural metadata production represents a reasonably high-effort obligation to meet in developing baseline services.

Administrative metadata

Administrative metadata for digitized text serve the same functions as for images: to document provenance, ownership, and rights of access to files; and to record technical metadata needed to fulfill present or anticipated future mandates for data packaging and long-term storage.

Digital image and text production

Scanning

Page images are the building blocks of choice for most genres of source textual materials in libraries: holograph manuscripts, typescripts, and machine-printed text. Images have the twin virtues of being relatively inexpensive to create—compared to costs of transcription (see "Text Conversion")—and of being highly amenable to repurposing: for display, printing, and processing, by optical character recognition (OCR) software to auto-generate searchable text.

The best scanners for text digitization are those that achieve the best balance of production (speed), quality, and handling for historic materials. Note that:

- **Production speeds** are more sensitive to bit depth than resolution settings. Document scanners tuned for high-resolution black-and-white (1-bit) imaging achieve higher throughput than their grayscale and color counterparts. In general, these scanners also are simpler to operate.
- **Quality evaluations** for baseline systems or services need only to map to the minimum requirements needed to support access. (For higher quality, see "Above Baseline Services for Consistent Quality.")
- **Handling** is the key criterion to assess scanners or scanning services for text digitization. Although there are meaningful differences in the native capabilities engineered into some scanners, local library policies are the criteria used to evaluate candidate systems, such as scanners or cameras and cradles.

Commonly-held bound materials, such as published books and serials (particularly those in library bindings) provide the best test of a library's policies for materials handling. Each of the approaches listed below has proven viable in production, as each is in current or recent use in library text digitization programs. Consider the following questions:

- May staff unbind volumes, then autofeed pages through a straight-feed scanner? If yes, then a Fujitsu M4097D scanner or its equivalent would provide *significantly* higher throughput than all other scanner types, particularly for two-sided pages since this duplex scanner digitizes both sides of a page in a single pass.
- If autofeeding of unbound pages is not allowed, may librarians disbind material? If yes, then flatbed scanners would be preferred, with technicians manually positioning each page on the platen.
- When bindings must be retained, may librarians to turn volumes over and to scan volumes face down? If yes, then flatbed scanners provide the most cost-effective solution.
- When bindings must be retained and materials must be digitized face-up, may librarians open books fully (180°)? If so, then four options are viable:

  (a) Use the hybrid approach: microfilm volumes, then digitize the microfilm.

  (b) Digitize the pages directly with an i2s DigiBook book scanner that automatically turns the pages.

  (c) Digitize pages directly with a conventional book scanner, where technicians turn pages manually, such as Zeutschel Omniscan or the Konica Minolta PS7000 scanner.

  (d) Digitize pages directly with a digital camera and a book cradle.
- When materials must be digitized face up, and bindings may only be opened partially (that is, less than 180°), then three options remain:

  (a) Use a Kirtas Technologies Apt BookScan™ 1200 robotic book scanner that automatically turns the pages.

  (b) Use a digital camera with two operators: one at the camera, the other holding the volume in position and turning the pages.

  (c) Use a digital camera; hold the volume in place with conventional exhibit cradles, or a custom book cradle or easel.

Fujitsu M4097D, www.fujitsu.com

i2s DigiBook, www.i2s-bookscanner.com

Zeutschel Omniscan, www.zeutschel.com

Konica Minolta PS7000, http://konicaminolta.us

Kirtas Technologies Apt BookScan™ 1200, www.kirtas-tech.com

Each scanner configuration introduces tradeoffs. None perfectly accommodates all bound volumes. Materials vary in size and condition; many include foldouts or other oversize inserts. Where book scanners include software optimized for page scanning, digital cameras have software optimized for graphics materials (that is, digitizing images).

High-end, high-production book scanners with robotic page-turning systems are priced in the six figures. To be cost effective, they are applicable only to high-volume operations.

When planning the configuration of baseline services for producing page images, librarians should consider one of two approaches:

- Commit to one method of production (and probably one scanner type) and outsourcing all other types of text digitization.
- If resources of space, staffing, and funding are available, invest in two or more scanner types to accommodate a reasonable range of text sources.

The Weissman Preservation Center, Harvard University Library, maintains a list of links to book scanners and easels, http://preserve.harvard.edu/resources/bookscanners_table.html

Text conversion

Electronic text (searchable text) can serve one or two purposes. Text may be needed solely to facilitate keyword searching of page images or to support keyword searching and to be displayed as a transcription—with or without annotations or other editing—that is quicker to display and easier to manipulate than page images.

Publications, such the Arts and Humanities Data Service's *Creating and Documenting Electronic Texts: A Guide to Good Practice,* provide an excellent overview to rationales and methods for text production.

The operations manager's concern is to determine whether baseline services must support production of hidden text for searching, displayed text for reading and other uses, or both. In general, the least expensive approach to generating hidden text is to create page images, use one or more OCR programs, then leave the OCR-generated ASCII uncorrected.

Text of high enough quality for display, however, either needs to be created by keyboarding (also known as keying), or by correcting OCR-generated errors by comparing OCR results to the source item. (In this case, source refers to source for keying, which might be original print, microform surrogate, or photocopy.)

See ITC links to OCR companies and products, http://tev.itc.it/OCR/Products.html.

OCR programs offer a range of capabilities—across document types, fonts, languages, and even handwriting—all to various degrees of accuracy according to the complexity of the source material and the quality of the page images fed into the OCR software.

Unless the entire work is to be displayed as a single file (for example, one HTML document), structural metadata are needed to put each ASCII file, corresponding to the original page, into the appropriate sequence. Adobe's portable document format (PDF) contains internal tags to structure searchable PDFs, as do programs that generate e-book formats from text masters.

Quality control

Choices of quality control metrics and methods for page images are identical to those for images produced in image digitization workflows. Claims of accurate color matching require above-the-baseline services.

For keyed or OCR-generated text, character accuracy is the conventional measure of quality. Although some OCR software generates confidence scores that predict the likelihood of errors being present on any page, *human comparison of the digitized output to the source, or manual keying of the entire page a second time (double keying) are the only reliable methods to ensure accuracy*.

Without double keying or manual comparisons of digitized text to the sources, librarians would be unreasonable to claim 100% accuracy to users, funders, or other stakeholders in a text digitization project.

Thus, many libraries choose to produce page images and OCR-generated text, and to make both formats available to users. The page images offer the data to the answer to the question, How accurate is the transcribed text?, but the user rather than a production technician invests the time making comparisons to discern the answer.

Data tracking, assembly, and packaging

Because digitized text *objects* are comprised of page-level units, text digitization yields many files. These must be assembled by software to build sustainable masters and functional delivery versions.

Fortunately, several of the open-source and commercial delivery products listed below (see “Delivery”) also bundle capabilities to package the component parts from each digitization workflow into coherent objects for storage and management.

Short- and long-term storage

Products of text digitization require the same storage infrastructure as those for images. (See Chapter 3 for details.)

Delivery

Standards for encoded text masters, including structural metadata objects, fall into one of three broad categories: SGML, XML, or PDF. In the case of PDF, the PDF/A (PDF-Archival) subset of the standard is recommended for master versions of multipage objects.

For SGML, the Text Encoding Initiative (TEI) standard has predominated; an XML migration is underway for TEI. Other XML flavors for standards-based production of text masters have been the Making of America II DTD, and, more recently, METS, the Metadata Encoding and Transmission Standard. The METS Implementation Registry documents the many uses of METS—not all of them in the text digitization domain—showing its rising popularity.

Which of these formats is best? This decision is a contextual one, influenced primarily by the delivery system that will be used for digitized text objects, and secondarily by the library’s longer-term interests in interoperability, where it is important to monitor policies and practices among some of the larger libraries managing digital repository services.

Delivery system requirements are fundamental to choices of image formats, image sizes, numbers of images per original work, and metadata formats. Although budgets to create and sustain these systems will likely be administered within the library’s digital library rather than digitization operation, delivery systems indirectly drive the budgets for metadata and digital imaging systems.

PDF-Archive, www.aiim.org/standards.asp?ID=25013

Text Encoding Initiative (TEI) , www.tei-c.org

Making of America II DTD, http://sunsite.berkeley.edu/MOA2

METS Implementation Registry, http://sunsite.berkeley.edu/mets/registry

As the most visible component of the baseline services of text digitization, delivery systems must fit the genres of sources and digitized text product types, user needs, budgets, and resident capabilities for programming.

The following brief list of front end products for digitized text collections conveys that systems range from home-grown and closed (Harvard PDS), to open source (Greenstone, NYU), to commercial and free (Acrobat Reader), and, finally, to expensive, but powerful products (DLXS and Olive) which bundle various middleware and back-end features with a robust interface and search engine.

Selected delivery applications

- Adobe Acrobat Reader, www.adobe.com
- Greenstone Digital Library Software, www.greenstone.org/cgi-bin/library
- Harvard University Library Page Delivery Service (PDS), http://hul.harvard.edu/ois/systems/pds/index.html
- New York University Library "METS Page Turner With Search" (XSLT-based page-turner and search implementations are freely available for use), http://dlib.nyu.edu/metstools/metssearch
- Olive Historical newspaper collection software, www.oclc.org/olive/default.htm
- University of Michigan Digital Library eXtension Software (DLXS), www.dlxs.org

Above-baseline services for consistent quality

The baseline services described above for text digitization are quality agnostic. With these services in place, a library could produce a digitized text collection, but not necessarily to any stated level of quality.

The following tables present some of the meaningful attributes of source, quality, and quality control likely to compel a program a program to invest in above-the-baseline services for text digitization.

Matrix of Attributes* for Above-baseline Layers of Service for Page Images

	Source type(s)	Digital quality	Quality control
Low effort	Unbound pages < 17"; no enumerated pages	Legibility	Check of completeness, pictorial attributes
Medium effort	Microform surrogates, illustrated items, meaningful color all permitted	Fidelity (dimensions and details); encoded pagination	Check of completeness, pictorial and digital attributes, pagination
High effort	Pages > 17"; bound material that must be scanned face-up; material in poor condition; illustrations; color; enumerated pages; meaningful hierarchical structure (many parts to work)	Fidelity + accurate tonal or color reproduction; encoded pagination and sections	Check of *matching* pictorial attributes; format validation; pagination; structure

* excluding descriptive metadata

Matrix of Attributes* for Above-baseline Layers of Service for Text and Encoded Text

	Source type(s)	Digital quality	Quality control
Low effort	Single digital format (such as all 1-bit, all 8-bit, all 24-bit page images), single language	Uncorrected OCR	Completeness (correct # of files), file naming
Medium effort	Multiple source formats or multiple languages; multiple layouts	Minimum *overall* accuracy threshold	Completeness + character accuracy (low % sample)
High effort	Poor or uneven quality of source pages (page images); multiple layouts; small type; meaningful parts (such as names) that must be encoded	Minimum accuracy threshold for each page or page component (such as authors and titles)	Completeness + character accuracy (statistically valid sample); validation of mark up

* excluding descriptive metadata

Low-effort strategies

Levels of service for text digitization fall into the baseline + low-effort category as follows:

- **Selection:** Efforts are made to constrain selection of sources to printed formats amenable to one text digitization workflow and one product type. All pages would be smaller than 17" in the long dimension. Sources selected to create fully searchable texts would be of a single language.
- **Materials preparation and handling:** Under certain constraints, published bound materials are permitted to be digitized intact on flatbed scanners or unbound to facilitate use of either flatbed or autofeed scanners.
- **Product choice:** Text digitization outputs are limited to page images only (of a single format), or to page images accompanied by hidden ASCII text for searching.
- **Scanning:** All digitizing is done by the library in-house (nothing is outsourced), without use of microfilm scanners, book scanners, or digital cameras.
- **Quality of page images:** The quality objective is to create legible reproductions.
- **Quality of text:** Accuracy requirements are limited to completeness—one ASCII file per page; no omissions of pages per work—without any prescribed thresholds of average (or minimum) accuracy per page. No requirement to display text. No explicit requirement to encode pagination for enumerated pages in sources.
- **Technical metadata:** Page images, text files, and structural metadata files would have checksums, as well as administrative metadata documenting ownership and rights.
- **Delivery:** Requirements for navigation would be modest: means to page forward and page back, and either to go to a specified page number or to go to a specified section for many-part objects.

Medium-effort strategies

Levels of service for text digitization fall into the baseline + medium-effort category when *any* of the following needs must be accommodated (even if only in one project):

- **Selection:** Source materials are selected for intellectual (content) value, regardless of their formats, age, dimensions, quality, and condition. Digitizing must fit the source.
- **Materials preparation and handling:** All bindings must be retained. Conservation review is mandated for certain types of material.
- **Product choice:** No constraints are imposed on number of product types for a collection. Workflows are established to produce different types of page images (such as bi-tonal for text pages, color for illustrations) within one multipage object.
- **Scanning:** Some or all digitizing is outsourced. Alternatively, one or more book scanners or digital cameras are purchased for the library's digitization operation.
- **Quality of page images:** The library purports to make *faithful reproductions*. The quality objective is raised from simply creating copies to creating copies that meet any pictorial criteria for goodness. In addition to the subjective methods described in the low-effort service configuration, technicians would be required to have sufficient visual literacy to compare copies with sources. (Objective methods of quality control, such as use of technical targets, would only be introduced if staff were appropriately trained.)
- **Quality of text:** Must satisfy requirement to support full-text searching via uncorrected OCR. Encoded text must be of sufficient quality to be displayed; enumerated pages in sources would be encoded. Statistically valid samples of pages would be inspected as complying with the specification for character accuracy.
- **Technical metadata:** In addition to checksums and administrative metadata, some preservation or technical metadata would be mandated in the workflow—whether stored internal or external to the image files—and verified to be accurate and complete.
- **Delivery:** Delivery systems would support all product types specified in the digitization workflow: page images, page images + hidden text, or encoded text (without page images). Printing would be supported (such as print page; print section; print entire object). Searching might be more granular than keyword.

At this level of service, a broader production infrastructure is in place, although the user would not necessarily have a choice of formats to access and display for any given work. Specifications of product types might vary within one collection or among collections.

Selection and preparation workflows would include a review component, whereby a technician searches one or more databases, such as the OCLC Registry of Digital Masters, to determine whether the work-in-hand had previously been digitized to an accepted level of goodness.

In using the Registry to record its intention to preserve its digitized text objects, the institution would either comply with the *DLF Benchmark for Faithful Digital Reproductions of Monographs and Serials* or comparable standard, or it would document and link to its locally developed specifications.

DLF Benchmark for Faithful Digital Reproductions of Monographs and Serials, www.diglib.org/standards/bmarkfin.htm

High-effort strategies

Levels of service for text digitization fall into the baseline + high-effort category when medium-effort capabilities are fully accommodated and *any* of the following needs are supported (even if only in one project):

- **Materials preparation and handling:** Bindings must be retained and volumes must be digitized face up.
- **Product choice:** Support for selection of sources and specifications for text encoding requiring domain expertise (such as scholarly analysis) for editing and mark up.
- **Quality of page images:** The library purports to make *faithful reproductions*, including management of color. Objective as well as subjective metrics and methods are used to verify quality.
- **Quality of text:** Minimum accuracy levels are mandated and verified for hidden text—either an overall average, minimum percentage per page, or minimum percentage per selected part (such as chapter titles). Must satisfy requirement to support full-text searching via uncorrected OCR.
- **Technical metadata:** In addition to checksums and administrative metadata, some preservation or technical metadata are mandated in the workflow—whether stored internal or external to the image files—and verified to be accurate and complete. Tools are used to validate formats.
- **Delivery:** Application provides means to search, navigate to different levels of a hierarchy for multipart works, and go-to-specified page numbers for enumerated pages. Printing supported. Middleware allows the user to switch views (to access different formats) and possibly to generate these deliverables on the fly.

All requirements for device calibration and color management noted in Chapter 3, “Image Digitization,” also apply to production of page images that must either accurately reproduce source materials, or meet a specification for pictorial rendering intent that depends on soft proofing.

Production of device-independent masters (see Chapter 3) also is highly desirable at this service level, although this is a more straightforward proposition for structural metadata and encoded text, where tools emphasize descriptive mark up to categorize parts of a document rather than to fix its appearance in any specific application.