Building A Digital Library: A Technology Manager’s Point of View
by Elizabeth J. Shaw

The Historic Pittsburgh project at the University of Pittsburgh can be instructive to others embarking upon digital library projects, and deciphering their own environmental opportunities and constraints.

From a case study perspective, the Historic Pittsburgh project illustrates the type of factors not uncommon when launching a digital library project: addressing user needs, determining content, evaluating the institutional infrastructure and available technology, and pursuing funding opportunities. This article focuses on how the technological and infrastructural environments at the University of Pittsburgh influenced the development of Historic Pittsburgh’s full-text collection by the Digital Research Library (DRL), a unit within the University Library System. It will examine how the staff of the DRL moved from a small pilot project to full-scale production while offering insight into some of the challenges that were faced. In conclusion the article will summarize some of the lessons learned in the course of the project.

DESCRIPTION OF THE PROJECT

The Historic Pittsburgh project (http://digital.library.pitt.edu/pittsburgh/) of the University Library System (ULS) at the University of Pittsburgh is a digital collection that provides an opportunity to explore and research the history of Pittsburgh and the surrounding Western Pennsylvania area on the Internet. It is a joint project to virtually gather the historic resources of the University of Pittsburgh and the Historical Society of Western Pennsylvania (HSWP). The Web site provides access to full-text collections, archival finding aids, real estate plat maps, and the Historical Society’s library catalog. Projects to provide 19th century census data and archival and photographic materials are under development.

Both the full-text and finding aid collections utilize Standard Generalized Markup Language (SGML) encoded text indexed by Open Text’s search engine and displayed on the Web using CGI “middleware” developed by the University of Michigan’s Digital Library Production Service. The full-text collection follows a model provided by the University of Michigan’s experience with its “Making of America” project. The real estate plat maps are compressed using MrSid software from LizardTech for viewing and selecting portions of the maps for zooming.

EXISTING INSTITUTIONAL INFRASTRUCTURE RESOURCES

The University Library System was an early experimenter with electronic texts encoded in SGML. Although most of the effort had faltered by 1997 due to staff turnover and other pressing IT priorities, an infrastructure for working with electronic texts remained. The University of Pittsburgh licensed Open Text’s SGML-aware search engine, a variety of SGML tools, and several SGML-encoded electronic texts, including the second edition of the Oxford English Dictionary and several Chadwyck-Healey databases. Library staff interested in learning SGML-encoding practices attended workshops. Further, UNIX systems administration expertise was available within the Information Systems department of the library. However, it was clear that additional technical support would be needed to develop and move any project beyond the most experimental phase.

In August 1997, the ULS demonstrated its institutional commitment to developing a realistic digital library project by appointing a working group headed by a full-time librarian whose sole mission was to explore options and approaches to developing a digital library project. Members of the working group included librarians, archivists, bibliographers, catalogers, and technical personnel. By the spring of 1998, a realistic assessment of...
opportunities and constraints emerged after some preliminary experimental work and hours of research into models and standards. As a result, the ULS established the Digital Research Library department and two major content priorities surfaced.

The University holds an extensive collection of 19th century schoolbooks that are widely used in a variety of fields of research, including English, cultural studies, and the history of education. The DRL recognized the important opportunity to study and disseminate this unique collection of materials electronically because many titles are rarely held in other institutions. In addition, there was strong interest expressed by members of the faculty, librarians, and school teachers in a project that would focus on the unique history of the region. The audience for this project would include social and industrial historians, students at the University and area high schools, genealogists, local community members, and a broad range of others. The project would also give the ULS an opportunity to partner with the local historical society whose rich content would complement its own. Thus, using the Internet to collate dispersed materials that existed in several different physical locations was particularly appealing to the project planners. Ultimately, the decision was made to concentrate first on what became known as the Historic Pittsburgh project. A project team was subsequently formed.

CHOOSING A DEMONSTRATION PROJECT THAT WOULD SUCCEED

Together the ULS and the HSWP hold thousands of monographs and serials, hundreds of archival collections, millions of photographs, maps, and historic documents relating to Western Pennsylvania history. Supporting the research needs of those studying regional history could entail digitizing a wide variety of historical materials in numerous formats, potentially creating a complex technical infrastructure. Moreover systems and software requirements for display and retrieval for these diverse formats vary significantly.

To simplify the first stages of the project, the team initially focused on one type that could integrate well technologically: published books out of copyright protection. The team recognized that investment in hardware, software, and technical expertise do constrain the choice of material that can be effectively digitized and made accessible.

The initial pilot for the Historic Pittsburgh project was funded from reallocated institutional resources with a view towards raising additional funding after demonstrating a successful project. Given the rich infrastructural support for SGML and associated technologies, an approach that incorporated these resources was not only viable, but lowered the entry cost by capitalizing on existing resources.

NON-CONTENT CONSIDERATIONS

After considerable research, the DRL identified several other non-content driven considerations that influenced its approach and overall philosophy towards the development of the Historic Pittsburgh project. These factors would influence the choice of materials to be digitized as well as the project implementation path.

- **Cost/Content Balance**—The DRL sought to provide maximum content provision at a reasonable cost.
- **Non-Proprietary Data Formats and Open Standards**—Given the rapid rate of change in emerging technologies, the DRL wanted to invest in resource data that could be shared across platforms and applications in non-proprietary formats. Using open standards and formats that are application independent would insure that data would not be lost to obsolescence of an individual software system.
- **Maximum Quality Data**—The DRL sought methods to create high quality data that matched the best practices and standards emerging in the digital library field. Data capture is very expensive, but the incremental cost of capturing high quality data are considerably less than the cost of rescanning or rekeying it again at a later time. The DRL also wished to reserve the option to reuse the data should better systems emerge in the future.
- **Availability of Metadata**—In addition to data capture, a large part of the cost of any digitization project is the collection and preparation of descriptive, structural, and administrative metadata. Finding a way to utilize descriptive metadata already captured in electronic form (e.g., MARC records) would greatly reduce costs and guarantee authenticity and authority of the works. Staff time would still be needed for the capture of additional structural metadata to describe the relationships between components of our digital objects, such as a table of contents. While examining the work of early implementers of digital libraries, the DRL tended to follow those that had developed some metadata standards rather than those where the issue had not been evaluated.
- **System Reliability and Scalability**—The DRL sought a model that would scale to reliably manage a large amount of data and that received broad-base support by other institutions involved in their own projects.
- **Added Functionality through Digitization**—The DRL sought a system that not only provided online readability, but also supplied additional functionality to the material. Capabilities to search the material, collate physically disparate materials, or provide new links between materials were important considerations.
- **Preservation of Originals**—If original materials were to be destroyed or damaged in the process of digitization, the DRL sought a means of providing a copy in an accepted preservation format.
- **Broad Access**—Although the primary audience might be campus users, a high priority was to enable access for the widest group of users. The DRL sought a model that used low bandwidth and required that the end-user need nothing more than a basic Web browser to access the resources.

CHOOSING THE FULL-TEXT COLLECTION AS A DEMONSTRATION PROJECT

The University and the HSWP maintain significant collections of 19th and early 20th century monographs and serials that are often used in conjunction, although the material is scattered throughout four University locations and the HSWP in downtown Pittsburgh. The ability to bring these materials together virtually for research and browsing on the Web, therefore, appealed to both institutions. The DRL debated the merits of delivering the full-text of the items versus the full-image display of the pages. The team at once recognized the full conversion of text to machine-readable form using double or triple keying was cost prohibitive. Furthermore, advisors to the project expressed concern that conversion to text-only, without the benefit of images of the original pages of the books, divorced the
books from some of the visual information provided in the original pages. Indeed, many of the books are rich sources of line drawings, photographs, and lithographs that considerably enhance the textual information. On the other hand, merely placing images of the pages online would allow remote users access to the materials, but it would not fundamentally change the level of access for our intended constituency—local users who could access the books at the various libraries in their current form. Neither image-only nor text-only solutions satisfied the DRL.

A unique set of circumstances converged to address this quandary. The University of Michigan had developed a hybrid approach to the digitization of historic resources with its creation of the “Making of America” (MOA) project. MOA provided access to 1,600 monographs and serials by presenting images of the original pages with underlying SGML-encoded “dirty” OCR to enable full-text searching. Data conversion for this model (scanning of images and OCR generation) is considerably less costly than keying the text. Although the uncorrected OCR provides imperfect searching, it does provide significantly greater access to the contents of a book than an image-only project can deliver. An opportunity to capitalize on Michigan’s efforts fell into the lap of the DRL when Michigan began to license its MOA “middleware” through the SGML Server Program (SSP), now the Digital Library eXtension Service, (DLXS). Furthermore, the DRL employed the author of this article, a former member of the MOA team, to lend technical expertise to the Historic Pittsburgh project. Once joining the SSP, the University of Pittsburgh’s DRL could begin to mount some of its existing SGML-encoded text collections using the Michigan “middleware” with nominal support from technical staff.

The MOA model also met most of the criteria that had been established as important non-content considerations:

- **Cost/Content Balance**—The hybrid image/OCR approach would get more material online at a lesser cost than full-text keying and SGML-encoding.
- **Non-Proprietary Data Formats and Open Standards**—None of the source data resides in proprietary format because the “middleware” uses SGML (ASCII text) and the widely adopted TIFF image format.
- **Best Practices**—Utilizing the recommendations of Anne Kenney and Stephen Chapman at Cornell University, page images for MOA were scanned as 600 dpi bi-tonal Group IV compressed TIFF images. This quality mirrored the best of image digitization practice, and it would allow for reuse of the images in the future.
- **Availability of Metadata**—Materials selected for the Historic Pittsburgh project had been described bibliographically in either the HSWP or ULS online catalogs, enabling the project team to capture and transform the descriptive MARC records easily. Issues of the document structure and file naming conventions had been well analyzed through work for the MOA project, providing a framework on which to build our own practices.
- **System Reliability and Scalability**—The system utilized for MOA had been proven to be reliable and scalable since it was managing 650,000 images. By comparison, the Historic Pittsburgh project scope was unlikely to exceed 250,000 images.
- **Added Functionality through Digitization**—By using OCR-encoded SGML as an underlying searching system, significant new functionality is gained by allowing users to pinpoint a book, a chapter or a page that contains the material of interest. Searches could also be performed on bibliographic data.
- **Preservation of Originals**—It was possible to generate acid-free facsimile reprints from the bi-tonal digital images and have these reprints bound and placed back into circulation. Since many of these materials were brittle, the reprint would extend the life of the print collection as a by-product of the digitization project.

**Broad Access**—The MOA model required a simple Web browser and provided data transmission at a relatively low bandwidth. Therefore, capitalizing on the existing MOA model would enable the DRL to move quickly to a demonstrable pilot project. The DRL receive strong support from the University Librarian to embark on this course.

**Expansion of the Model**

Although the MOA project met the criteria that the DRL had specified from both a technical and philosophical perspective, there were some important Web interface aspects lacking which the DRL staff and project team felt were important to address. For example, there were only minimal links to major document structures provided in the MOA interface. To add a hyperlinked table of contents, the DRL modified the metadata collection to add document structures, such as chapter breaks and illustration headings.

**Preparation of the Pilot**

Having determined the model and identified the modifications that would realistically enhance the project within the institutional constraints, the DRL staff worked to develop a pilot project. The pilot project was designed not only to test the extensibility of the MOA model to the University of Pittsburgh environment, but also to develop workflow and production processes that would be required in implementing a full-scale project. A history bibliographer at the ULS identified 20 significant books in the public domain. An analysis of the widely varying document structures of the late 19th and early 20th century books was performed, and information about document structure and pagination were captured in a spreadsheet during collation of the books. The pilot images were scanned locally on a flatbed Hewlett Packard scanner at 600 dpi or on a Minolta planetary scanner at 400 dpi. Optical character recognition (OCR) was subsequently performed on the resulting images after each image was checked to ensure that it accurately represented the page. Descriptive bibliographic information was extracted from the MARC record and transformed to an SGML-encoded header for the digital edition of the work. Technical staff experimented with and modified the MOA “middleware” to present a navigable table of contents. Most importantly, the pilot project gave staff an opportunity to identify points in the process that might effectively be automated.

By September 1998, a modest pilot project was ready for demonstration to a strong local library advocate, the Hillman Foundation. The demonstration garnered enough support and funding to proceed with the first production phase of the Historic Pittsburgh project.

**From Pilot Project to Production**

Although the pilot demonstrated the technological feasibility of our approach, sev-
ereral decisions about the production process needed to be made. In the pilot phase of the project, many tasks were performed on a book-by-book basis. Moving into a production mode to digitize 500 books over the course of two to three years required that staff identify processes that could be automated or assisted by automation as well as those where human intervention would dramatically increase the quality of the product. Through several iterations the DRL adapted a workflow that balanced the need for speed, accuracy, automation, and human intervention. Moreover, the DRL staff grew from a coordinator and technical projects manager to include a production librarian and assistants.

Metadata Capture/Collation of Materials

The DRL decided to expend the greatest amount of staff effort in the earliest part of the production process to provide the information for a navigable table of contents and automated SGML-encoding. It takes hours of work to collate a book page-by-page, record significant document structure, transcribe significant headings, systematically assign unique filenames for each page, write instructions to the vendor for special handling of particular pages, and assign codes for automatic SGML page encoding. This work was later guided by a detailed procedural document, which reached final form only after experience with at least 100 books. This early, yet critical, investment for every book smoothed the way for many automated production steps and provided value-added information to the online product.

Scanning

Digitizing over 500 books by the DRL would have required a heavy investment in scanning equipment, staff supervision and training, development of a production workflow, and correction of quality problems, not to mention securing additional workspace in an already overcrowded facility. Thus, the DRL secured a vendor that could handle all aspects of the high-volume imaging process and could better manage imaging costs and production flow. The identification of the vendor and the Request for Proposal process was greatly aided by the RLG Tools for Digital Imaging. Working from their sample RFP and modifying to the particular circumstances at the DRL, competitive bids from vendors were secured and a vendor selected based on its ability to meet imaging and facsimile reproduction needs.

The DRL project staff would be responsible for the preparation of materials and checking the quality of materials returned from the vendor.

Quality Review

The process of checking the quality of the vendor’s work played an important role in the success of the project. The DRL staff examined both the facsimile reprints created to replace the originals and the digital files themselves for problems such as skew, poor image quality, and incorrect orientation or page order. Initially every page in the facsimile reprint and every digital image were checked for accuracy. However, based on the quality of the work from the vendor, the quality checks were later reduced to a complete check of the facsimile reprint, checks on all significant document structure (e.g., first pages of chapters, illustrations), and a 25 to 30% check of unspecified text pages. This process is assisted by scripted tools that use the metadata spreadsheet to identify and present the page images for specified document structures and a random sampling of the plain text pages.

Optical Character Recognition (OCR) of Text

Although the DRL purchased software to perform batch OCR, significant improvement in the quality of the OCR could be obtained by using pre-zoning tools on pages that have illustrations—particularly on those that had illustrations nested within text on a page. Evaluating each and every page is time intensive and costly. Again, utilizing the document structure information from the metadata spreadsheet, another script presents to the staff only those images that will benefit from pre-zoning. Zoning selected pages reduces the error rate on unusual pages and improves the speed with which the batch OCR process runs. The remaining images are OCR’ed with automated batch processes. Because 19th and early 20th century books have multiple typefaces, unusual or broken type faces, poor quality images due to poor quality originals, and skew in the original published material, these factors have reduced the optimal OCR capture rates of 99%, but still manage to provide high rates of retrieval in the online system.

Preparation of Materials for Online Access

The final stage of the production process involves preparing the raw materials—bibliographic information about the original text, structural metadata, OCR’ed text and images—for online display. Although the online system displays images of the pages, the underlying navigation and searching mechanism is provided by searchable SGML-encoded text. The structural information in the metadata spreadsheet, bibliographic information from the MARC record encoded in SGML, and the OCR’ed text provide the data for an automated encoding process.

To make viewing and navigation of the online books easier, the project team decided that illustrations with non-upright orientations should be rotated to a comfortable viewing orientation and that original pages that were blank should be identified as such. Another script rotates and/or inserts targets on a batch basis since this information was collected during the initial metadata collection stage. Once the SGML is indexed by Open Text and the images placed on the server, the full-text collection is ready for public display.

Lessons Learned

The University of Pittsburgh’s Digital Research Library faced a unique set of opportunities and constraints when developing the Historic Pittsburgh project. What lessons can be shared?

Consistent and Complete Structural Metadata

The project team initially decided to collect structural metadata about the books to increase access for users and to make it easy to encode the SGML automatically. The metadata spreadsheet also quickly became a fundamental tool in automating some of the more tedious and time consuming production steps. The spreadsheet was used to automate aspects of the quality control process, enhance the OCR through pre-zoning, rotate images, and encode the SGML because it contains information about unusual book structure and notes images requiring special attention. Although the initial effort to capture the metadata is high, the time spent has been more than saved by this automation. Collecting the data at the initial stage of collation has also assisted in ensuring that the resulting digital images and facsimile reprints are an accurate representation of the original books.

Close Collaboration between DRL Staff

One of the strengths of the DRL staff has been its ability to communicate closely with
each other. Communication about how the online system should work helped to influence the data capture methods and design of production processes. Moreover, communication about problems with production steps allowed technical staff to develop automated routines to handle the most tedious tasks. When processes were modified in the course of streamlining, production staff could then anticipate the impact of changes on other processes.

**Cost Tradeoffs**

The DRL staff believes that, although the introduction of significant metadata collection and pre-zoning of unusual images for OCR introduced a complexity into the workflow, this effort was warranted for the increased access that it provides to the material. This has not, in fact, been demonstrated objectively and warrants further study; in the future, the DRL hopes to study the validity of these and other cost/quality tradeoffs.

**Mid-Course Corrections**

The DRL was forced to develop production processes as it went along since it was a relatively early implementer of a local digital library of published texts. This has on more than one occasion required that staff reexamine and redesign a process to make it more efficient. Flexibility in rethinking the problem has been crucial.

**Access and Display Tradeoffs**

The model chosen does not display complicated grayscale or photographic images very well since the source images are bi-tonal. However, the tradeoff is worth it since it is easy to create facsimile reprints from the images, the images can be handled in a consistent manner, and the capture of bi-tonal images is comparatively inexpensive. As a result, there are books that have not been chosen for the full-text collection precisely because their primary or significant informational content is contained within photographs or illustrations that would best be captured as grayscale images. This limitation has subtly influenced the scope of the collection and will be addressed in future modifications of the project.

**Building on Others’ Experience**

The DRL has developed a project providing access to monographs and serials that builds on the work of others while addressing local needs and desires. The DRL took considerable effort to seek out and evaluate the work of digital library pioneers, such as the Library of Congress’ National Digital Library Program, the University of Virginia Electronic Text Center, and the Berkeley Digital Library SunSITE. This effort was well spent since the DRL could learn from the challenges and pitfalls faced by other institutions. By paying attention to the work of others, it became possible to move forward more quickly than if we had tried to develop this expertise in-house.

**Conclusion**

The path taken to develop the full-text collection of the Historic Pittsburgh project is uniquely its own, even as the staff of the DRL sought to follow standards and best practices established by experts in the digital library field. While the Historic Pittsburgh Web site has been more heavily utilized than expected, serious evaluation of the use of the full-text collection is an important next step in expanding on this initial success. Studies are underway to understand comparative retrieval rates on OCR’ed and keyed text from 19th and early 20th century books. Production processes are being evaluated to determine where other steps can be effectively automated and where human intervention can significantly increase the product quality. Capitalizing on the unique resources available at the University of Pittsburgh, even as there are fiscal, staff, and infrastructure constraints, has allowed the DRL to mount a widely used significant digital resource.

**Notes and References**

1. At the time there was considerably less information available about emerging standards and best practices, but the working group closely followed the work of Anne Kenney at Cornell and read the materials provided by the Library of Congress for its National Digital Library Program.
2. See http://www.umdl.umich.edu/moa/.
4. More information about the University of Michigan’s DLXS program for SGML-encoded materials is available at http://www.umdl.umich.edu/dlxs/.
7. Initially pages with reports that the OCR software had low confidence in its reporting of characters were flagged and minor corrections on personal names, place names, and significant words were performed. This has turned out to be extremely time consuming and tedious, and, therefore, costly. It has also been difficult to establish clear guidelines to instruct staff members when to correct or skip the ASCII text. Complete correction would be even more costly. After running comparisons on retrieval results between data that has had this step of manual intervention and that for which the OCR process is fully automated, insignificant improvements in retrieval rates were obtained. The results of this analysis will be published in a future article.

**Resources for Those Interested in Starting Digital Libraries**

Listed below are only a few resources for getting started. Many of them contain links or references to many other valuable resources.


**Web Sites That Point to or Contain Resources for the Planning and Implementation of Digital Projects:**