Many More than a Million: Building the digital environment for the age of abundance

Report of a One-Day Seminar on Promoting Digital Scholarship
Sponsored by the Council on Library and Information Resources

November 28, 2007

Final Report

March 1, 2008

Table of Contents

Context

The Council on Library and Information Resources (CLIR) hosted a workshop in Washington, D.C., on November 28, 2007, to talk about the general problem of what can be done with the very large digital collections now taking shape as a result of mass digitization projects, the so-called "million books" problem. This was the third of what will ultimately be five workshops held on this topic, organized by Tufts University with funding from The Andrew W. Mellon Foundation. The first took place in November 2006 and focused primarily on the issues

surrounding classical studies. A second more general meeting that looked at a range of humanities subjects took place at Tufts University in May 2007. Subsequent meetings, aiming at a European audience, are planned for London and Berlin in March 2008. This sequence of workshops on the million books problem converges with CLIR's programs in digital scholarship, cyberinfrastructure and preservation and is part of an extended, distributed conversation on these related topics that CLIR is supporting in several venues in 2007 and 2008.

The November CLIR meeting was particularly important in that it included not only humanists, librarians, and computer scientists but also representatives of the National Endowment for the Humanities (NEH), the National Science Foundation (NSF), the Institute of Museum and Library Services (IMLS), the Library of Congress, The Andrew W. Mellon Foundation, and Google, a list which illustrates the enlarged participation that digitization and computationally assisted research have created in the world of humanities scholarship. Although the discussion focused on text collections, we considered four major questions that are largely format independent but illuminate the implications of scale:

- What is the problem? How does access to large corpora of digital materials change that problem?
- What services do scholars need?
- How do we manage digital collections when the digital material is abundant rather than selective?
- What systems or infrastructure is necessary to provide services and materials to scholars?

The problem

Two and a half millennia ago, Plato in the *Phaedrus* critiqued the static nature of writing: the written word can no more answer questions that we may pose than can the painting of a human being. Digital services now allow us to begin redressing the static nature of recorded information. We possess a growing array of digital instruments with which to represent and analyze the world: geographic information systems allow us to

analyze space at a large scale, 3D modeling systems allow us to model objects and buildings, and visualization systems allow us to detect patterns in large data sets. Increasingly we produce structured data: DNA and protein sequences; data collected by sensors in the oceans, satellites above us, instruments in supercolliders; and so on. It is easy to see the explosion in the volume of information. Less obvious is the transformation in the notion of content, whether it is called *data*, *information*, *objects*, or *documents*.

Until now, libraries have been able to work with books, in the sense that books are containers of ordered content that conforms to a logic. In a digital world, users want the information inside the document, which may be variously known as "content" or perhaps as the "logic" as distinct from the physical artifact. They do not want the page but rather particular units of information—a dictionary entry that may begin on one page and end on another, a single piece of information that may be stored as propositional rather than textual data, or a data set that appears as a table spread over many pages. Machines read digital collections (or subsets of them) for us, converse among themselves, and in some instances can begin to construct answers to our questions. Where clay tablets, stone inscriptions and papyrus are silent, our documents have begun to find their own voice: a digital translation of Thucydides' *History of Peloponnesian War* can, in effect, ask a named entity identification system to present to the curious reader information about the particular Alcibiades who plays a prominent role in the history (as opposed to the many other figures with that name); a digital article in German can call for a machine translation system to give speakers of English an initial idea of what the article says.

But if we have new methods with which to structure and interact with knowledge, what do we do with thousands of years of human knowledge and expression that is available only in print form? We can scan books by the million, but print culture produced static expository text, tabular data, hand-drawn maps, and reproductions of photographs, all designed for human readers with far deeper understanding of the world and of the print conventions than any machine can at present attain. Even human readers can only decode a subset of the content within written documents. As Plato points out in the *Phaedrus*, the words "iron" or "silver" each carry meanings that are commonly understood. Terms such as "just" and "good" are much less well defined. To modern eyes, Plato

has given an example of "disambiguation," that is, associating a term with well-understood references, whether the word "iron" describes a metal or an action.

As we shift the written record of all human culture into the digital environment and thus—for the first time—potentially into the center of human intellectual life, we will have only enough human labor to convert writing on stone, clay, pottery, papyrus, vellum and other physical media into image form. We cannot retype even the ten million books in the Harvard library system, much less read and annotate the contents of each page with significant XML markup. And even if we could marshal the resources to do so, the human life contains only about 30,000 days: reading a book a day we would finish a million books only after thirty lifetimes of reading. Only machines can process or "read" the vast written record of humanity.

No longer a distant probability, a digital representation of this written record is taking shape before us. Google Books, the Open Content Alliance, the Million Book library and other major digitization efforts all depend upon the image front searching technique popularized in the 1990s by projects such as JSTOR[1] and the Making of America.[2] This technique involved scanning the physical item to create digital images of pages and then subjecting those pages to optical character recognition (OCR) systems. At the time, these systems were considered sub-optimal because of the incidence of errors and the output was familiarly called "dirty OCR." But when the dirty OCR was paired with the page images, the value of the technique became greatly enhanced because the text was considered good enough for searching, and potentially ambiguous results could be compared with the page image, which was considered authoritative. This strategy of linking page images with OCR enables us to make effective use of large corpora of relatively cheaply scanned books and was, in large measure, effective because it points backwards to the limitations of print: search gets human readers to the page and leaves them to parse out its meaning.

But what happens when landing at the page is not sufficient? If a reader lands in a page image of Latin or Old Norse, the reader had better have a

[1] http://www.jstor.org/

[2] http://quod.lib.umich.edu/m/moagrp/

good background in those languages because the image front library will provide little if any help understanding what the page means once it appears. If a reader encounters a document about an archaeological site, the reader will need to recognize and match locations of catalogued objects to the print-based site plan (assuming that the scanning project was able to manage fold-out maps and other non-standard features). If a reader encounters a diagram of an object, the reader will need the skills to understand the conventions of that diagram.

In short, digitization does provide scale (or quantity) but does so at the price of rich, largely manual encoding. Visualization, customization, personalization, and similar analytical services increasingly familiar to us depend upon born-digital objects in which a great deal of structural and semantic knowledge has been encoded. The information captured on page images is, by contrast, implicit and often not directly accessible to the machines that will be always their first, often their only, and arguably their most important readers. So given the trade-off between scale and encoding in converted text corpora, what are our options, and where are the opportunities?

Why multi-million book libraries are different

The digital collections that emerged over the first generation of development are relatively small, with large digital collections generally containing tens of thousands of books. Mass digitization projects such as Google Book Search and the Open Content Alliance are creating collections in the hundreds of thousands and millions—collections large enough to begin modeling the scale and complexity of real academic libraries.

Very large collections based on image books differ from first-generation digital collections by one or more orders of magnitude. Of course, they are much larger: one internal estimate of Google's collection in spring 2007 suggested that their searches were scanning at least 2 million books already. At the same time, these collections are much more heterogeneous, with books from any library shelf likely to find their way into the scanning workflow. The range of subjects is thus far broader than in the curated collections to which we are accustomed. The range of subjects means that error rates will be much more variable, with OCR of many texts in non-standard scripts (e.g., Arabic, Classical Greek)

producing little or no searchable text. While good book-level metadata exists for many books that find their way onto library shelves, users of image books work with the contents of individual pages, e.g., chapters and sections, and content that includes proper names, technical terms, and text quoted from other sources. Pioneering services have begun to identify references to proper names in full text but the heterogeneity of content raises challenges here as well. The Washingtons mentioned in a document composed in 1780 are very different from those that appear a century later when more than a hundred cities in the United States had taken Washington as their name.

Large digital collections, many created as a result of mass-digitization projects, enable us to investigate new research topics. Such topics may cross several traditional disciplines or require a scale of data not available in existing collections. Examples of potential new research topics include the following:

- Linguistics. Automatically track patterns in morphology, syntax, and semantics across large stretches of time, space, and culture. These studies might be synchronic (e.g., comparison of American, British, and Indian English) as well as diachronic (e.g., the development of English over time).

- Intellectual history. Dan Cohen, assistant professor in the Department of History and Art History at George Mason University, pointed out that the secularization thesis, for example, states that the role of religion declined in general discourse during the 19th century, but most studies of this topic have been anecdotal. If we could track references to the Bible or to other religious terminology across thousands of texts, we could begin to put this thesis on more solid footing. Other topics might include analysis of the poetry cited in magazines, newspapers, or other popular literature, or the changing role of Shakespeare as evidenced by the plays mentioned and passages quoted.

  The notion of a collection organized by theme or research topic, such as the secularization hypothesis, is one strategy for building collections. Jonathan Bengston, associate librarian for scholarly resources at the University of Toronto, offered a second model: relational intellectual history applied to an individual's body of work.

For example, the John M. Kelly Library at the St. Michael's College in the University of Toronto is coordinating an effort to digitize the works of John Henry Newman, including monographs in all editions, sermons, newspaper articles, and manuscripts. The digitized text is being fed into a database and software is being used to identify subtle changes in language and meaning over time. Once the conversion has been completed (a task that will likely lead to a substantial updating of Newman's bibliography), the team will apply this software to Newman's corpus. Over time, it will be interesting to feed other materials that Newman would have been aware of into this database and see whether relationships can be traced between the evolution of his thought and the wider intellectual milieu at a far more granular level than has been achieved by traditional scholarly methods.

- Social and economic history. Will Thomas, professor of history at the University of Nebraska-Lincoln, pointed out the implications of very large collections for the study of broad socioeconomic topics such as the influence of railroads in U.S. history. Researchers can mine very large collections for references to, and propositional statements about, railroads to trace their development during the 19th century. Such records are also an important resource for understanding migration and settlement patterns; they enable regional comparisons within the United States (for example, the Plains versus the Central Valley of California), across borders (for example, the United States versus Canada), and over time (U.S. economic development in the 19th century compared with that of emerging nations in the 20th century).

- Cross-linguistic and cultural studies. Many complex, world historic topics (e.g., the rise of Islam) involve far more languages than any individual scholar can fully master. Very large collections contain the dictionaries, encyclopedias, parallel source texts/translations, gazetteers, and similar print reference works from which machine-actionable knowledge can be mined. Researchers working with cross-language information retrieval, translation support tools such as machine-readable dictionaries and morphological analyzers, and even imperfect machine translation can work with a broader range of linguistic materials than was

possible before. We can already begin to use text in languages such as Arabic and Chinese.

What services are needed?

First-generation image-front digital libraries such as JSTOR and the Making of America allow users to search for words in the so-called "dirty OCR" automatically derived from image books and then view the original scanned page images. Such services reflect the practices of print culture: in effect, we enhance access to print books by adding search capability and allowing readers to transmit pages in digital form, but, in the end, we are delivering conventional printed pages. LCD displays have simply joined clay tablets, stone, vellum, papyrus, parchment and paper for the display of static written text. Although our habits of reading have changed throughout the history of static writing,[3] these changes build on the assumptions of the relationship between the human reader and static text.

The exemplary tasks described above and the increasing prominence of interdisciplinary work both demand more complex services. During the workshop, Wendy Pradt Lougee reported results from an Andrew W. Mellon Foundation-funded survey at the University of Minnesota (UMN). She observed that humanists, like their colleagues in the big sciences, now find themselves working with materials from more disciplines than they can readily master, and the ability to understand the conventions of multiple disciplines constrains the questions that scholars can now pose.[4] Consider an excerpt from one of the UMN faculty interviews:

> My manuscript is actually quite interdisciplinary[5] in the sense that I've written many parts of it with the understanding that I'm talking to other scholars outside of history, outside of my

---

[3] Kathleen Fitzpatrick, "CommentPress: New (Social) Structures for New (Networked) Texts," Journal of Electronic Publishing 10.3 (2007): http://hdl.handle.net/2027/spo.3336451.0010.305.

[4] Wendy Pradt Lougee, A Multi-dimensional Framework for Academic Support; Final Report submitted to the Andrew W. Mellon Foundation from the University of Minnesota Libraries, June 2006; http://www.lib.umn.edu/about/mellon/UMN_Multi-dimensional_Framework_Final_Report.pdf.

[5] Ibid., p. 21

> own trained disciplinary field. Talking to American Studies scholars, or sociologists, African American Studies as a department here is an interdisciplinary department. So although I'm an historian, I interact on a daily basis with sociologists, psychologists, humanists, American Studies, Cultural Studies. (Faculty interview, November 2005).[5]

This faculty member faces two fundamental challenges: understanding and then communicating with intellectual communities that have different assumed background knowledge, different ideas of what questions are and are not important, different conventions of argumentation, and so on.

The interdisciplinary challenges this assistant professor faced are daunting enough. For topics requiring the use of multiple languages, the challenges are much greater and have prevented in-depth work. A topic may involve a manageable amount of source material (for example, a few thousand pages) but that scale means little if the sources are in a dozen languages or if they are not translated. Even if they are translated, readers must have knowledge of the culture from which the sources emerged to understand their cultural and intellectual context.

Ultimately, researchers need services that can customize and personalize the materials with which they work, providing just-in-time intellectual support that matches the background and the momentary intentions of a given user at a given place. Three computer scientists working with the DARPA-sponsored Global Autonomous Language Exploitation program[6] identify core processes needed to make large bodies of analog data useful: document understanding, multilingual services, and converting raw text into machine-actionable data.

> **Document understanding (Thomas Breuel, DFKI):** Page images need to be converted into machine-actionable data. At the simplest level, this includes OCR-generated text to support searching. Users have, however, more demanding queries: e.g., locate translations of book 1, chapter 86 of Thucydides' *History of Peloponnesian War* in English, French, and German, as well as all commentaries and dictionary entries that comment on particular word usages. Such a

[6] http://projects.ldc.upenn.edu/gale/.

query assumes that we can not only convert print to text but also isolate logical chunks of text (e.g., the particular section of Thucydides, or the articles of a dictionary with their headwords). We need to provide not only accurate keystrokes but also semantic markup.

Multi-lingual services (David Smith, Johns Hopkins): We need to be able to convert as much information as we can from one language to another. This includes not only machine translation but cross-language information retrieval: classicists are supposed to be proficient in (at least) Greek, Latin, English, French, Italian, and German and usually prefer to view documents in their original language but would benefit from being able to pose a query in one language and then search the other five languages. Often, users may have some knowledge of a language and seek translation support tools. This would include the user with a year or two of language study as well as the expert viewing text from a new domain (e.g., a classicist reading a Renaissance Latin text, where every word was known from the classical period but the words have acquired new senses in the Renaissance community that produced that text).

Converting raw text into machine-actionable data (David Mimno, University of Massachusetts at Amherst): This includes various first-order classification tasks; for example:

- Know that *fecit* is the third-person single perfect indicative form of the Latin verb *facio* (morphological analysis);
- Recognize that Washington in a given context describes a place and also know which Washington is being referred to (e.g., Washington State, Washington, D.C., or one of dozens of places called Washington—georeference);
- Distinguish references to George from Booker T. Washington (general named entity recognition);

- Not only recognize that "Th. 1.32" is a citation to a particular chunk of a canonical text, but also be able to distinguish where it describes book 1, chapter 32 of Thucydides vs. line 32 of Theocritus' first Idyll or line 32 of Aristophanes *Thesmophoriazousai*.

Based on these first-order classification tasks, we want to be able to build more complex structures: namely, recognize not only that *agricola* (Lat. "farmer") is nominative rather than ablative in a given context, but also that it is also the subject of the verb *fecit* ("the farmer did…"); or convert "Caesar in Alexandria" into "Julius-Caesar present-at Alexandria-in-Egypt." We also want to be able to mine significant patterns that may or may not be associated with particular named entities or classification schemes: for example, what topics occur as words or phrases in a particular document clusters.

The CLIR workshop provoked spirited discussion about who would provide what services. Joyce Ray of IMLS acknowledged that the institutional repository systems in universities have not developed as quickly as many had hoped.[7] She suggested that institutional repositories could grow more rapidly if they concentrate on archiving and preserving the electronic files that the institutions are obligated to maintain, such as electronic theses and dissertations, instead of depending on faculty members to deposit publications voluntarily. She also said that there is a continued need for traditional archival values, a point that resonated with many participants who had observed problems with versioning and quality control in the initial collections that the mass digitization projects have released.[8]

Okan Kolak from Google gave a presentation on Google Book Search, describing the project's advances in quality control and its increasing range of services. Google already supports the mapping of place names automatically discovered in the full text of a scanned book—a service, Crane noted, that Perseus had developed for its digital library in the late 1990s but that no library had

[7] See, for example, Rieh, Soo Young et al., "Census of Institutional Repositories in the U.S.," D-Lib Magazine, November/December 2007: http://www.dlib.org/dlib/november07/rieh/11rieh.html.

[8] See, for example, Paul Duguid, Inheritance and Loss? A Brief Survey of Google Books. *First Monday* 12 (August 2007), http://www.firstmonday.org/issues/issue12_8/duguid/.

stepped forward to support in its repository. Google was also hard at work on many of the other services that researchers at Tufts had identified as key to intellectual labor, such as general named entity identification, document structure analysis, and automatic quotation identification. Crane observed that there is cause for concern: While libraries have slowly developed repositories focused primarily on basic delivery of simple objects such as preprints and images, Google would become the default supplier of those services that exercised commanding value. He challenged the room to articulate what services the academic library community would provide and warned that, given current conditions, disciplines would build their domain-specific services on top of those offered by corporate giants such as Google and Microsoft, resulting in unanticipated dependencies down the road.

Collections

Thomas Garnett, of the Smithsonian Institution, plays a leading role in the Biodiversity Heritage Library. He led the argument that, while players such as Google, Microsoft, and Yahoo might offer useful services, many in the academic world and beyond had serious reservations about entrusting the long-term preservation of our cultural heritage to commercial entities, however enlightened or far-sighted their current positions. Academic libraries and national institutions such as the Smithsonian Institution and the Library of Congress have maintained our historical and scientific record over long periods of time and the public good is central to their missions.

But our own research institutions and time-honored practices also create barriers. Recalling his own recent research for a presentation on the Gettysburg Address, Will Thomas listed the sources he used including American Memory, Valley of the Shadow, ProQuest Historical Newspapers and other digitized newspaper collections, his own private sources, and those that colleagues had recommended. Like a 19th century traveler shifting from stagecoach to canal boat to railroad car, Thomas, one of the principal authors of the Valley of the Shadow project, had to examine one information silo after another to explore a basic event in American history.

Thomas hopes that his project on railroads will prove a good way to explore the challenge of integrating historical data from different sources and different formats (e.g., digitized newspapers, history engines, statistical data housed by the Inter-university Consortium for Political and Social Research), since railroads were networks that connected cultures and languages with multiple sources of data (land grants, drawings, texts, images) and illustrate the simultaneity of growth across space and time. Thomas hopes that all of this data can be shared and lead to "open research environments," although he experienced difficulties with his library in trying to create a truly open research environment. For example, while the researchers were able to access a server, they had trouble getting permission to experiment. Computer scientists are well aware of the tension between servers set aside for production and operations, which must be stable and robust, and those dedicated to research, where the goal may be to interrupt the system precisely to reveal its limitations. But supporting such experimentation is a new role for libraries, which are keenly aware of their service mission to the university and their infrastructure role on campus.

Whereas Garnett articulated an ideal world in which cultural heritage institutions would provide access to our cultural heritage, Thomas pointed out that, in the first generation of digital projects, a fragmented information world has emerged, where libraries have, in large measure, financed the creation of information silos, many in the hands of corporate interests. Libraries that are members of the Association of Research Libraries (ARL) invest about $1 billion per year in their collections[9] and could, alone, have undertaken a long-term project to digitize millions of books. Instead, organizations such as Carnegie Mellon University's Million Book Library, the Internet Archive, Google, Yahoo, and Microsoft have undertaken the grand challenge of converting the published record of humanity into digital form. Individual libraries and library systems have contributed. The University of Michigan (UM) stirred controversy by

[9] "ARL member libraries make up a large portion of the academic and research library marketplace, spending more than $1 billion every year on library materials." (http://www.arl.org/arl/index.shtml). Specifically, as of 2005/2006, the most recent data publicly available from ARL, total library materials expenditures reported by 113 university libraries is $1,159,553,716. Total library materials expenditures reported by 10 non-university libraries in the same period is $66,804,538. See M. Kyrillidou et al., *ARL Statistics 2005-2006* (Washington, D.C.: Association of Research Libraries, 2008), p. 42.

partnering with Google, but the Michigan/Google agreement gives Michigan rights to make non-commercial use of its holdings and prohibits Google from charging fees for searching books digitized from the UM libraries.[10] The University of Toronto and the University of California library systems were among the early partners who have allowed the Internet Archive to launch the Open Content Alliance.

In this context, Ray made several significant observations:

- Her agency has begun to see (and to fund) efforts within individual disciplines that can lay the foundations for cross-disciplinary data curation. For example, IMLS is supporting a project at UCLA to move its Cuneiform Digital Library from a university department into the university's digital library system, where it will be archived with appropriate metadata and maintained for long-term access and preservation. In another project, library staff and astronomers at Johns Hopkins University are working with publishers, the National Virtual Observatory, and colleagues at the University of Washington and the University of Edinburgh to establish a digital library framework and archiving principles for content that can be used in publishing research in astronomy. These kinds of projects will enable digital library data curators to establish models that can be applied in other fields. This will promote interdisciplinary research and help avoid the creation of discipline-specific data silos. A cautionary tale demonstrating the need for such cross-disciplinary efforts can be found in the case of the Alexandria Digital Library (ADL), an early digital library of geographic data. The ADL, developed at the University of California at Santa Barbara in the 1990s, provided searching primarily by latitude and longitude, as these data are universally known by the geographers for whom the library was designed. However, this made the data inaccessible to

[10] "4.3 Searching Free to the Public: Google agrees that to the extent that it or its successors make Digitized Available Content searchable via the Internet, it shall provide an interface for both searching and a display of search results that shall have no direct cost to end users. Violations of this subsection, 4.3, not cured within thirty days of notification by U of M shall terminate U of M's obligations under section 4.4." The Google/Michigan agreement is available at http://www.google-watch.org/foia/umfoia.html as well as at http://www.lib.umich.edu/mdp/umgooglecooperativeagreement.html.

non-geographers, and even the geographers found retrieval increasingly difficult as the library grew in scale.

- Universities cannot afford to support multiple preservation repositories. University libraries, because of their mission, expertise, and historical experience, are often the best positioned institutionally to support long-term preservation of and access to digital assets.

- The emerging field of data curation is making a significant impact on professional education in library and information science (LIS). This will create a cadre of specialists with the necessary skills to manage large-scale digital libraries. IMLS's 21st Century Librarian program—begun in 2003 to address the projected retirement of many librarians—has enabled LIS schools to move beyond preparing replacement librarians, and to focus on identifying and addressing the skills and competencies that librarians and archivists will need to support the massive digital libraries of the future. IMLS has funded several projects to develop programs in digital asset management and data curation. Data curation programs have been established recently at The University of North Carolina at Chapel Hill, the University of Illinois at Urbana Champaign, and the University of Arizona, and others are being developed. Evidence suggests that a data curation curriculum will be deployed rapidly throughout LIS education.

The strongest emerging models seem to combine massive scale with the flexibility for particular domains to manage data and provide services that suit their needs. In such models, generic services link vast multidisciplinary collections with specialized services providing more advanced access for particular questions. The Internet itself represents the historic example of such a collection. Google Book Search and the Open Content Alliance are, by contrast, relatively modest efforts, but they reflect the best balance between curated collections (which require some centralized attention and resources for each unit added) and completely open collections, which are not centralized, controlled, or otherwise restricted.

Data conversion remains the greatest challenge. Even if we leave aside rights issues, we have no realistic framework within which to convert the published record into digital form. Ray noted that the discussion of infrastructure should include the important questions of how to find the right economic models and the right economies of scale for both large-scale digitization and for preservation repositories. As of February 2008, the corporate entities that have funded the largest digitization projects had not made any explicit commitments to long-term preservation. At the same time, there is a danger that the historical record will be distorted if only books and other printed resources are available in digital form. There is not strong funding support for digitizing much of the rich heterogeneous content held by museums, archives, and library special collections.

Even if we have in place the right business models and economies of scale, we must decide what we need to collect and how we want to collect it. In the print world, the library was done once it had purchased a print volume, cataloged it, and placed the physical copy somewhere where it could be retrieved. Collecting page images of a book is only the start of a process that can include careful transcription and potentially open-ended markup. In one model, based on the Text Encoding Initiative, there are five levels of markup.[11] The simplest level is image books, where we have only a TEI header for metadata and uncorrected OCR-generated text. The fifth level includes scholarly editors' emendations about textual variants; linguistic features such as morphology, syntax, and word sense; proper nouns; and other categories of markup.

Humanists working with very large collections need at least four different types of data conversion.

- **Raw OCR output from page images with human-curated book-level metadata** is the starting point for all work. Automatically generated metadata, capturing structural elements (e.g., tables of contents, chapters/sections, footnotes, marginalia, indices) will be sufficient for many purposes.

[11] Friedland, L., N. Kushigian, et al. (July 30, 1999). "TEI Text Encoding in Libraries: Draft Guidelines for Best Encoding Practices (Version 1.0)." Retrieved May 26, 2000, from http://www.indiana.edu/~letrs/tei.

- **Curated structural metadata.** If we have carefully marked the headwords of a reference work such as the Encyclopedia Britannica 13th edition, we can then use the OCR-generated text for each article to train classifiers to distinguish references to various Springfields or Washingtons in unstructured text. Likewise, scholars may find it useful to augment automatic text alignment by marking the book and chapter boundaries in translations of a canonical work (e.g., English translations of Livy's *History of Rome*). In this case, we focus on accurate structural metadata but allow subsequent generations of automated systems (including OCR and named entity analysis systems) or community-driven efforts to correct and provide deeper markup for the text. Other examples include analytical cataloging for magazines and journals, and catalogs of objects from museums or archaeological sites. For example, the widely adopted Smithsonian Trinomial system is built on a single alphanumeric designation that identifies each officially recorded archeological site by state, county, and site.[12]

- **Curated transcriptions.** Canonical literary works such as the Greek text of Homer's *Iliad* and the plays of Shakespeare are objects of intense study and their readers have little tolerance for uncorrected transcriptions. Such texts will demand careful production and then a mechanism by which to correct any residual errors. It may be enough to have accurate transcriptions of one or several editions and to use these carefully produced transcriptions as a framework against which to align and collate dozens or even hundreds of other editions for which only OCR-generated text is available.

- **Structured data sources.** While encyclopedia entries about people, places, organizations, and other topics often comprise loosely structured expository text, some print reference works are proto-knowledge sources: they strive to represent structured data in a regular format. Decades ago, researchers working with the Oxford English Dictionary discovered that human editors working with human authors for a human audience do not create regularly structured data—they always unconsciously depend upon the

[12] For example, see Texas Archeological Research Laboratory, Site Records, http://www.utexas.edu/research/tarl/records/site.php.

intelligence of their readers to see past inconsistencies and to fill in missing data. Extracting machine-actionable data from such sources has typically required substantial investment: accurate keyboarding and then customized software to parse the typographic conventions of the print original into usable data. The results can, however, repay the investment. Print indices that distinguish one Washington from another in a given context provide training data for machine-learning-based systems that can analyze millions of books. Scholarly lexica can contain foundational morphological and syntactic data, as well as hundreds of thousands of citations that associate particular passages (referenced with canonical citation schemes) with particular word senses.

Cost is a major consideration in any research project, and particularly so in complex, multi-language projects, issues of context and disambiguation have become more acute. Where scanning image books may cost about 10 cents per page, keyboarding a large dictionary page with Greek and English may cost $10—two orders of magnitude more. Developing dictionary/index/gazetteer readers that can be customized to extract high-value, domain-specific information is a messy problem. The most important data are often those that do not lend themselves to general approaches. One example is the ability to recognize the highly abbreviated canonical text citations on which much humanities scholarship depends (e.g., where "Th. 1.38" describes book 1, chapter 38 of Thucydides' History of the Peloponnesian War, line 38 in the first Idyll of Theocritus, or something else). Nevertheless, a mature collection development policy for very large collections should probably include not only massive scanning but the production of curated knowledge bases. While page images may be sufficient for the kinds of materials that would circulate in a print library, we may want to convert the reference room into a highly structured knowledge base that supports human readers and automated systems alike.

Systems

Although we may have a model for the services we want, a handful of tools, code to run them, and a substantial collection of data, we may not have the systems in place to do what we want where we want to do it. The experience of Thomas and his colleagues is indicative. The issue they

faced was not one of tools but of a suitable environment in which to experiment.

Consider the Open Content Alliance (OCA), which offers more than 300,000 books for download, with a particular emphasis on publications from 19th century North America. This constitutes, potentially, the most important new collection ever available for the study of the culture and history of that period. We could use it to explore in new ways topics such as changes in religious language in scientific discourse or the development of railroads. No users, however, seem to have in place the systems to work with even the 300,000 volume collection already in the OCA. For some purposes, downloading the OCR-generated text from the OCA would be enough, but we will often need to run our own OCR. The OCA, for example, does not provide page break information in its OCR and we cannot thus go from the OCR output for a given document to the page image from which it was derived. More importantly, if we need to tune OCR for a particular domain (e.g., Latin vs. English) or for a domain in which conventional OCR produces no useful data at all (e.g., classical Greek), we may need to run the OCR on large segments of this corpus again—that is, assuming we can identify in some general way subsets relevant to a particular topic.

Google has far more content online than does OCA and has the resources with which to analyze that content, but it is unclear how Google could open up those computational resources to a more general public. Okan Kolak posed a number of questions to consider:

- What to expose (image, text, relations, citations, quotes, keywords)?
- How to expose it (fixed collections, internal access, APIs)?
- What are the legal, technical, commercial, social considerations?

But even if Google and others continue to provide better content, we are still left with managing the data deluge and the challenges that deluge create for libraries. José-Marie Griffiths, dean of the School of Library and Information Sciences at the University of North Carolina at Chapel Hill, talked about these issues. Vast collections of digitized books are just one category of data that libraries need to manage. In addition are the scientific data sets that have become objects of persistent value and now find their way into the library. With such resources measured in petabytes of data, millions of digitized books may become relatively minor issues.

Libraries need to plan strategically for four categories of infrastructural services that provide the foundations on which the more advanced user services described above must depend. They are core components of a comprehensive cyberinfrastructure:

1) network connectivity
2) computational capacity
3) information discovery, integration, and analysis capabilities
4) intra-and inter organizational fluidity and "sanity"

Griffiths laid out a model of cyberinfrastructure comprising a hierarchy of levels. The base technologies include computation, storage, and communication. Networking, operation systems, and middleware are at the next level, followed by a level of essential services such as high-performance computing systems, data and information management, observation measurement, interfaces and visualization, and collaboration services. At the top level of cyberinfrastructure are the higher-order services such as community-specific knowledge environments, research portals and gateways, and customization for discipline- and project-specific applications.

Griffiths believes that the library has the most significant role at the level of higher-order services. Like Lougee and Ray, Griffiths stressed the librarians' function in mediating between general and domain-specific services. With the growing amount of data already available online, there is a major need for tools to enable simultaneous, seamless searching across multiple information resources, such as library catalogs, digital libraries and databases, and Web-based resources, as well as integration tools to ensure that researchers can find, organize, manipulate, and analyze relevant information from many data sources. Libraries alone have traditionally dealt with all disciplines and have an important role to play in building systems that support collaboration across disciplines.

It is unclear how existing library infrastructure would build the systems on which scholarship will depend. While libraries have a long history of collaboration in developing and sharing metadata, a cyberinfrastructure will require much deeper collaboration than anything that we have seen before.

Major Questions

The history of ideas in the West is replete with examples of great questions that unified disparate strands of research: the mind/body problem in philosophy, the relationship between magnetism and electricity in the history of science, the sources and duration of the Renaissance in Italy and Northern Europe, the rise of the middle class in Europe, and the significance of the frontier in American history. And there are others. The great questions can emerge over time, enabling successive generations of investigators to see how their work builds upon, revises, or even sets aside work that preceded it. Or they may be articulated relatively quickly, like the grand challenge questions in science. Such “marquee questions,” a term Clifford Lynch used, provide large-scale coherence. As Griffiths has noted, articulating broad, driving topics helps the research community leverage resources efficiently. Possibly the most significant finding of the workshop concerned the importance of articulating “marquee” questions and ideas about how to identify them. Timothy Tangherlini of University of California at Los Angeles (UCLA) suggested that a registry of projects might be maintained in which investigators identified the questions that they sought to address, enabling relationships among various projects to become apparent.

So, this workshop posed one big question: How does scale in content, made possible by mass digitization, change humanities research? As a result of the discussion, we can now pose five questions that parse this broad question into more manageable topics for research:

(1) How do traditional archival values migrate into the computationally intensive environment made possible by copious digital data and digital tools? The digital "text" or object becomes plastic, in the sense that it may be devoid of context, may be modified itself (e.g., by a spell checker) or by the addition of automatically generated markup (e.g., named entity identification or machine translation), and may be displayed differently in different systems even if the “content” remains unchanged. Thus, there is a need for ways to establish authenticity, provenance, and integrity of digital sources as well as versioning, which is a known problem in the mass digitization collections.

(2) When only the computer actually "reads" the object or the text, a new and not fully understood relationship is created among author, tools,

objects, and readers (or users). Traditional paleography and criticism address such relationships among written and printed material documents; how do we model and understand the digital equivalent? Presentation of analog source material in digital form still involves mediation whether by editors, coders, or machines. What is the shape and form of that mediation?

(3) What happens when large-scale, team research becomes possible and perhaps even necessary, enabling interdisciplinary research? We have already begun to see dynamic, interactive online productions, like the *Valley of the Shadow*,[13] the *Walt Whitman Archive*,[14] the *Dickinson Electronic Archives,*[15] and *Uncle Tom's Cabin and American Culture,*[16] or a future contribution to an archival database like the one under construction at the University of Toronto for the works of Cardinal Newman. Will we see much larger projects covering topics and corpora more analogous to the grand publication series of scholarship (e.g., the Patrologia Latina) or even microfilm projects (e.g., an edition of 19th century newspapers combining machine intelligence and human training data)? The Perseus Digital Library was conceived in the 1980s as a critical mass of information about the ancient Greek world and went beyond any single author, genre, medium, or time period for standard projects. In a world of open content and interoperable systems, this critical mass already has started a larger process of collaborative production, partially visible in the Stoa in general (www.stoa.org) and the Demos project on Athenian democracy in particular (www.stoa.org/demos/). Lougee pointed out that the term "interdisciplinary" takes on different meanings in different disciplines. To humanists, it has meant reading the literature outside of the core journals in the traditional disciplines rather than interaction in work groups. There are several issues to resolve, among them:

- Attribution of authorship and credit, necessary for both promotion and tenure reviews and for funding requests to key public and private agencies.

[13] http://valley.vcdh.virginia.edu/
[14] http://www.whitmanarchive.org
[15] http://www.emilydickinson.org
[16] http://www.iath.virginia.edu/utc/

- Recognition of the value of the digital research and its expression in digital form, even if that expression is subject to ongoing change by subsequent generations of scholars. Not only should, for example, the digital Valley of the Shadow project be the authoritative work, but it should be seen as dynamic rather than a fixed, complex object so that collaborations take place over time as well as across traditional disciplinary boundaries.

- Recognition of digital scholarship that focuses on infrastructure. Classicists spent centuries creating the critical editions, lexica, grammars, encyclopedias, commentaries, and technical studies on which twentieth-century scholarship largely depended. We now need machine-actionable knowledge bases that can serve advanced systems and human researchers alike. These knowledge bases have print antecedents but the need to represent them in machine-actionable form and to support complex services may ultimately render them qualitatively different from their print predecessors. We need in the humanities as in the sciences to attract and support some of our most promising scholars to bridge the gap between the needs, present and potential, of the humanities and the possibilities enabled by scholarship.

(4) What are the infrastructure requirements? What belongs to the national cyberinfrastructure that is made available locally? What is maintained centrally on campus? What functions and affordances are appropriate at the desk top? And where are the dependencies? Griffiths identified the core components of a comprehensive cyberinfrastructure.

- network connectivity
- computational capacity
- information discovery, integration, and analysis capabilities
- intra-and inter organizational fluidity and “sanity”

She then offered a four-level model, showing how the infrastructure systems articulated with local and disciplinary needs. Every institution will face choices concerning what it will support, but arriving at those decisions can benefit from ongoing dialog with other communities within the campus, the respective disciplines, and higher education at the sector level. Sorely needed will be a framework within which such decisions can

be made to ensure a locally appropriate level of service, system-wide efficiencies, and sufficient redundancy to protect critical resources.

(5) Finally, what are the big questions—what the group dubbed the "marquee" questions—that justify the expenditure that managing digital information requires, that can be pursued when scholars have access to massive amounts of data, and that leverage the individual efforts across institutional and generational boundaries? Rarely are such questions articulated in a single meeting; they become apparent as investigators experiment, share results, and continue the discussion. As one participant commented, a “boutique project” may actually be an “experiment.” A niche interest or “boutique project”, which might be considered a singular topic, becomes an experiment, or a test of a large proposition, when the research is put in a larger intellectual context, by the author or by those who read, critique, and build upon successive research projects. Such coherent, overarching themes and broad implications may not always be obvious to the investigators themselves. But awareness of the notion of intellectual context is critical, and Tangherlini’s bottom-up suggestion for a registry that makes visible the self-organization of ideas is a creative step toward monitoring and capturing thoughts that might otherwise be distributed in workshop reports, conference proceedings, blogs, journal articles, and so on. For now, it is important merely to acknowledge the importance of questions.

Recommendations

We may be reaching a limit for what can be accomplished by discussion alone. We do not simply have a set of exemplary research questions for which million book collections are well suited. We can point to a set of services (for example, named entity identification, quotation identification, and mashups from multiple data sources) that users clearly want and for which open source code bases exist. Our library community increasingly understands that data curation will be a central library function and that such stewardship includes not only the data but also the services we use to analyze and augment that data. So to borrow a practice from our colleagues in computer science and engineering, we recommend starting to prototype some ideas to see how they work and to use the results of these experiences to inform the ongoing discussion.

We propose the following topics as priorities for future work:

1) **Finding ways to provide analytical access the Open Content book data now available should be a priority.** Scholars should be able to pose questions that analyze very large collections: e.g., what passages from Shakespeare or the Bible appear in different genres over time? What sorts of things are said about railroads in 19th century writing? Find all the Latin and translations of Latin authors that appear in documents over time. Apply specialized OCR to find a particular class of text scattered throughout the collection (e.g., classical Greek). Large collections of image books are a core data set for humanists and straightforward applications of high-performance computing could make a rapid impact on some fundamental classes of questions for many humanists.

2) **We should apply exemplary questions to open collections such as the OCA, access functions that Google, Microsoft, and others provide to end-user services and APIs.** Some of the examples presented by Okan Kolak about advanced searching raised questions among workshop participants. It was not clear that the results, though superficially impressive, yielded information of real value. Google and its audience would benefit from systematic feedback from the users who understand their domains best. Informal communications with individuals at Google suggest that the company would be receptive to building such an API but that they need a coherent, well-organized response from humanities scholars.[17] Developing requirements (at least in a general sense) and providing feedback and evaluation may warrant substantial resources.

3) **We need to clarify the costs and benefits of book scanning vs. the intensive transcription and markup of complex knowledge sources.** We cannot afford to apply human labor and expertise directly to more than a tiny percentage of the published record of humanity. Are there printed materials that would, if carefully converted into machine actionable form, uniquely enhance our ability to analyze relatively unbounded bodies of

[17] Daniel Cohen, personal communication, February 14, 2008.

material? Extracting the morphological data from large lexica of Greek and Latin was expensive and laborious, but limited. Once we had completed this task, we were able to apply heavily used searching and reading support services to open-ended bodies of Latin and Greek. We need to move toward general guidelines for more nuanced collection strategies that combine massive digitization with careful conversion of print into machine-actionable knowledge.

4) **We need to understand more clearly how to relate high-value, domain-specific services and data structures to services and data structures that are common to all collections.** Every discipline needs text searches, but some communities need different kinds of search. For highly inflected languages such as Greek and Latin, users need to be able to pose a single query and retrieve hundreds or even thousands of forms (e.g., ask for Latin *facio*, "to make/do," and retrieve *fecit*, *factus*, and *facis* but not *facilis*, "easy," *factionibus*, "factions" etc.). This implies that our underlying system architecture has a general slot for language-specific morphological analyzers. The morphological analysis problem thus reflects a generic function for which many solutions may be produced (e.g., morphological analyzers for Arabic, Russian, Greek, or Sanskrit). Small disciplines, such as classics, need to be built on the most general system possible and to focus on their own domain-specific problems. Thus, they should spend their labor producing the Greek and Latin morphological analyzers that would work with a general system that could support searching and analyzing the highly inflected languages on which their research depends. Likewise, different communities may find that they need to fine-tune even the most generic services: many communities want to associate proper names with their referents (e.g., Washington as person or place, city, or state) and language-independent methods may provide good initial results, but each domain will want to make sure that the general system can address its particular needs. Historians of science are not, by and large, specialists in Greek, Latin, and Arabic, but they need to be able to adapt systems for literary texts in these languages so that they can recognize the technical terms, idioms, proper nouns, and other features in texts about scientific topics that may have been created a thousand years after the last Roman emperor.

Finally, we should examine the future education and training required to create a new information professional. We are looking for a new type of information manager as well as a new type of scholar. Like the scholar of the future, the information manager of the future will be well-versed in the relevant technologies, capable of adapting to a changing environment, and able to anticipate the diverse challenges of both research and the pedagogical mission of the university. Some resources will be locally housed yet made available more broadly; others will be obtained elsewhere but will appear local to the end user. Managing the infrastructure to support a seamless dialog between the local and global will fall to the library of the future. As the Boomer Generation ages and is replaced by Gen X, Y, and beyond, identifying, recruiting, and training professionals, and nurturing their career development so they will be ready to staff that future place remains our most important challenge.

Acknowledgements

This report was written by Gregory Crane and Amy Friedlander with the superb assistance of Alison Babeu, who was rapporteur for the day-long session. We are grateful for the presentations by the speakers, several of which are presented in full elsewhere, and for the enthusiastic and thoughtful contributions of the participants during the meeting and the subsequent review of this document. We would also like to acknowledge the assistance of Crit Stuart, Program Director for Research, Teaching, and Learning at the Association of Research Libraries.

List of Participants

Jonathan Bengtson
Associate Librarian for Scholarly Resources
University of Toronto
Robarts Library, Room 6044/6045
Toronto Ontario M5S 1A5
Canada
Coordinator, University of Toronto/Internet Archive Scanning Centre

Brett Bobley
Chief Information Officer
National Endowment for the Humanities
Office of Information Resource Management Room 203
1100 Pennsylvania NW
Washington, DC 20506

Alison Babeu
Research Coordinator
Perseus Project
Tufts University
Medford, MA 02155

Hugh A. Cayless
Head, Research and Development Group
Digital Library at Carolina
Adjunct Faculty, UNC School of Information and Library Science
Wilson Library, CB #3990
UNC Chapel Hill
Chapel Hill, NC 27514

Daniel Cohen
Assistant Professor
Department of History & Art History
George Mason University
Robinson B 359
4400 University Drive, MSN 3G1
Fairfax, VA 22030
 Director of Research Projects
Center for History and New Media

Gregory Crane
Professor of Classics
Tufts University
Medford, MA 02155
Winnick Family Chair of Technology and Entrepreneurship
Editor-in-Chief, Perseus Project

Tim DiLauro
Johns Hopkins University
Digital Research and Curation Center

Rachel Frick (Address current 10/31/07; Ms. Frick moving to Institute of Museum and Library Services)
Head, Bibliographic Access & Digital Services
University of Richmond
Richmond, VA

Amy Friedlander
Director of Programs
Council on Library and Information Resources
1755 Massachusetts Avenue, N.W. Suite 500
Washington, DC 20036-2124

Thomas Garnett
Biodiversity Heritage Library
Program Director
Smithsonian Institution Libraries Rm. 22
MRC 154 PO Box 37012
Washington, D.C. 20013-7012

Stephen M. Griffin
Program Director
Division of Information and Intelligent Systems (IIS) National Science Foundation
4201 Wilson Boulevard, Room 1125
Arlington, VA 22230

José-Marie Griffiths
Dean, School of Information and Library Science
University of North Carolina at Chapel Hill
100 Manning Hall, CB #3360
Chapel Hill, NC 27599-3360

Charles Henry
President
Council on Library and Information Resources
1755 Massachusetts Avenue, N.W. Suite 500
Washington, DC 20036-2124

John B. Horrigan
Associate Director
Pew Internet Project
Washington, DC

Okan Kolak
Research Scientist

Google Book Search.
1600 Amphitheatre Pkwy, Mountain View, CA 94040

Wendy Pradt Lougee
University Librarian, McKnight Presidential Professor
University of Minnesota

Clifford Lynch
Executive Director
Coalition for Networked Information
21 DuPont Circle NW
Washington DC 20036

Kathryn Mendenhall
Director, Partnerships and Outreach Programs Library Services
Library of Congress
Washington, D.C. 20540-4900

David Mimno
Research Assistant
University of Massachusetts
140 Governors Drive
Amherst, MA 01003-9264

James J. O'Donnell
Professor of Classics
Provost
Georgetown University
37th and 0 Streets NW

Washington, DC 20057

Joyce Ray
Associate Deputy Director for Library Services
Institute of Museum and Library Services
1800 M Street, 9th Floor
Washington, DC 20036

Robert Rynasiewicz
Professor
Department of Philosophy
Johns Hopkins University
Baltimore, MD 21218

Roger Schonfeld
Manager of Research
Ithaka
151 East 61st Street
New York, New York 10021

David A. Smith
PhD Candidate
Computer Science Department
Johns Hopkins University
3400 N. Charles St.
Baltimore, MD 21218

Timothy Tangherlini
Professor / Head
The Scandinavian Section, UCLA
Box 951537
Los Angeles CA 90095-1537

William G. Thomas, III
John and Catherine Angle Professor in the Humanities
Department of History
University of Nebraska-Lincoln
615 Oldfather Hall
Lincoln, NE 68588

Andrew J. Torget

Director
Digital Scholarship Lab
University of Richmond
Project Director
Virginia Center for Digital History
University of Virginia

Donald J. Waters
Program Officer, Scholarly Communications
The Andrew W. Mellon Foundation
140 East 62nd Street, New York, NY 10065
Email:djw@mellon.org