When mass digitization reaches critical mass:
Scholars' evaluation and analysis of major digitization projects

> Extensive and reusable digital collections are at the core of humanities and social science cyberinfrastructure. Scholars must be engaged in the development of these collections.
>
> The extensive digitization of cultural heritage materials is one of the most exciting developments in the humanities and social sciences in the past century, and it should be continued and expanded through a thoughtful combination of institutional, public, and private support. The Commission believes that scholars have an important role to play in the development of commercial and non-profit digital archives alike, and neither research libraries nor companies such as Google have yet gone far enough to encourage dialogue with the scholarly community on such questions as the selection of materials for digitization, decisions about to omit from the digitized representation, or the design of descriptive metadata.
>
> *Our Cultural Commonwealth. The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences.* From Recommendation 8, p. 38.

Background

The long boom in what we thought of as mass-digitization projects of the 1990s exploded into something an order of magnitude larger with the arrival of the Google Print project in 2004, and its sequelae, the Open Content Alliance, and Microsoft's Windows Live Book Search, all of which remind us as well of the antecedent Project Gutenberg. Those projects hover like the visiting alien ships in *Independence Day* over a mass of other smaller projects, many of which have critical scholarly value. We are now either at or very close to the point, however, where the body of originally analogue material now in digital form is of such quantity and quality that we must facilitate the operationalization of broad-scale distributed digital libraries. Looking closely at the quality and functionality of these projects for *scholarly* users is vital to making sure that operationalization supports these vital social uses, ones that Google and others are not likely to have as first priority.

Most of the scholarship and discussions of large digitization projects touch upon three themes: intellectual property; technical aspects; and more abstract musings on the potential social impact of the project, often referencing the 'common good.' Copyright issues pervade the publications focusing on these very large digitization projects. The contention between publishers, for instance, and the Google project, which has resulted in several high profile legal challenges, some including publishers that have agreed to give content to the project, has been widely aired in the media and the blogosphere. The copyright complexities are compounded by the range of agreements Google has established with the major universities involved with the book scanning project: some are selectively giving Google books only our of copyright and in the public domain; others, like Michigan, are allowing the wholesale scanning of its entire holdings regardless of the book's copyright status.

The technical aspects of these large digitization projects have also received a great deal of scrutiny. Discussion has focused on technology itself, rather than the overall utility it affords, though issues that may be characterized as quality control are linked to discussions of technology. Google's refusal to make public its search algorithms, and Microsoft's perceived proprietary approach to almost all of its products, are often cited as problematic. The lack of cooperative planning in the design of the technical platform for these projects, which precludes interoperability and federation of the data, is another point of concern: "massive amounts of disorganized information," in Jean-Noel Jeanneney's terms. The quality of the digital images of the scanned books has generated blogs and other fora that detail the many representational problems of the projects. Pictures of scanned book pages with the scanner's fingers apparent, as well as misspellings and lost text from books held at odd angles to the camera, are prevalent. Web sites report the ongoing discovery of lacunae within selected works, as well as instances of illegibility.

In what might be termed the sociology of mass digitization, observers alternately contend that large-scale projects such as Google's and Microsoft's will enable new discovery of literary and other works not currently accessible to the public, will democratize knowledge, and will contribute to the public good in unprecedented ways. Others fear that the information held in these projects will be eventually sold as a commodity, decreasing access to it for the less affluent. Given the extraordinary costs associated with these projects, there may not emerge any competition to Google or Microsoft, and the market will be thus tightly held as a near monopoly.

These issues and the variety of interpretations they encourage are evidence of the complexity, and newness, of vast digitization projects. Yet almost invariably missing from the observations and evaluations of these mass digitization efforts are the perspective and voice of the scholar. The potential for scholarly research that the digitization of millions of volumes of books may entail is not difficult to intuit. The transformative nature of having such a sweeping array of our cultural heritage accessible, and that can be queried, interpreted, and reconstituted as new knowledge is, in abstract, astonishing. The issues attending technology, social impact, and legality inform but do not illuminate the core problem of mass digitization from a scholar's point of view: in order to support research these digital databases must be organized and architected to best reflect the methodologies and intellectual strategies of contemporary scholarship. The near absence of participation by scholars in the design and execution of projects such as Google Book Search and Microsoft's Open Content Alliance almost assures that the needs of this community will not be addressed.

This project is undertaken with four main goals in mind: 1. to assess selected large scale digitization programs by exploring their efficacy and utility for conducting scholarship in a variety of disciplines; 2. to write a report based on that research that summarizes the results of the project and makes recommendations for improving the design of the mass digitization projects, as well as recommendation for future development; 3. to convene a meeting of prominent scholars after they have reviewed the findings of the report in order to engage in more programmatic discussion concerning recommendations, assumptions, and next steps; 4. to create a Collegium of scholars that can serve over the long term as an advisory group that can collaborate with the corporations involved in mass digitization to help assure the highest quality of data and the highest level of utility; 5. to convene appropriate stakeholders from scholarly, library, publishing, and digitizing communities to discuss ways and means of effecting appropriate improvements..

Process of Evaluation

An informal categorization of the criteria that working scholars can use to structure their investigation and assess the usefulness of these projects follows. On many of these criteria it will appear that current projects fall short, not through any intrinsic weakness, but through the inability on all sides to recognize the impact that they can have when genuinely integrated, interoperable, and intermingled with traditional materials and new born-digital materials as well. To identify these criteria and begin the conversation about how to meet them is to begin to think appropriately big enough for the scale of these projects and the value they can bring.

Outline of Analysis

This project will focus on Google Book Search, Microsoft's Live Book Search, Project Gutenberg, Perseus, and the ACLS Humanities E-Book project as the main sources for analysis of mass digitization efforts. (To be determined: whether the Open Content Alliance has a body of sufficiently coherent material available to make this kind of analysis useful. If not, we may need at least to mention in report the challenges of a large project that does not proceed with the consistency of content and format that, e.g., a proprietary project charging subscription fee can do.)

The project will be conducted along the following outline.

1. Summary of main classes of projects and a few representatives of each. The project will employ two to three scholars drawn as opportunity provides from two to three distinct disciplines, e.g. classics; Renaissance history; and modern American literature. This selection of fields reflects our expectation to engage scholars from historical and literary areas of study whose interests are more readily supported by the datasets under scrutiny. This assures a range of chronological expertise and scope, and also considerably reduces the time needed to become familiarized with and adept in using the digital resources. Reluctantly omitted for purposes of avoiding scope creep and facilitating an expeditious report are disciplines drawing heavily on visual materials contained in printed volumes. The scholars will note their background and research interests, and comment briefly on important methodological considerations in the conduct of research as it pertains to their discipline.

2. General statement of elements working humanities scholars will look for in such resources. While there will be unique aspects pertaining to the individual disciplinary approaches, some general characteristics of inquiries against a large database often include:
   a. ease of discovery in an integrated interface (OPAC or better)
   b. ease of access to full text
   c. ability to copy, quote, cite, search
   d. ability to share results of work with other scholars and students

e. ability to reprocess the digitized text in order to support other kinds of research and use – "Web 2.0 functionality"
f. predictability and reliability of the "collection"
g. differentiation by field of types of material contained in "mass digitization" projects of interest – scholarly books, primary source material (novels, memoirs, documents), materials not published in print form, and periodical literature (scholarly and non-scholarly)

3. "Core samples" of documentation
The critical goal of this project is for the investigators to perform a defined routine of inspection of sample product from each of the main projects under scrutiny and report upon type and frequency of quality issues to be found. The scholars will execute their work following a template for assessment (see below).

The PI and consultant will collaborate to outline a set of search strategies designed to test different kinds of books (differentiated by discipline, language, and format). There will be an initial period of working with the sampling staff to determine likely rates of productivity and thus ultimate sampling sizes possible. Core attention will go to Google Print (50% or more of sampling), 25% to other broad-brush digitization (Microsoft, OCA), 25% to specialty projects (e.g., old Chadwick-Healy digitizations). These percentages may be adjusted after consultation with the advisory committee.

Participating scholars will summarize their assessment in individual reports. The reports will be synthesized, with prioritized recommendation extrapolated from them. While major concerns will be cited, emphasis will also be given to exemplary projects that go beyond the norm in providing some of the desiderata and meeting some of the concerns above.

The resulting summary will be promulgated and form the basis of a larger meeting of scholars to discuss and to determine the most efficacious means of moving forward, including ways to work with individuals and corporations developing the databases to assure that responsiveness to scholarly methods and desired applications and tools will be factored in the design and future architecting of these projects.

Criteria: A Template for Assessment
Scholars will respond to the following criteria as part of their evaluation of the databases, noting specific areas of concern in these terms with digitization projects.

1. Availability of MARC record

2. Access to full text
    a. pre-1923 materials (presumptively out of copyright)
    b. post-1923 issues (presumptively in copyright)

3. Quality of interface
    a. ability to download a useful full text
    b. ability to print selections or full text

c. ability to retain a working copy on the scholar's desktop or in some other form of user-organized filing system (e.g., through use of a program like Zotero, or through a provider's system, e.g., similar to Google Picasa for storage and organization of images)
d. ability to cross-search texts from different projects and origins, hosted on different servers

4. Quality control issues
a. sample percentage of scans defective for scholarly use
b. quality of scan and usability for reading, searching, printing extracts, print on demand in book form, or OCR
c. completeness of scanned volumes (front matter, back matter, footnotes, endnotes, bibliographies)
d. overall taxonomy of scans (from Juliet Sutherland of Distributed Proofreaders, who also maintains a log of bad Google scans) – poor/reference/recognition/archival/reproduction

5. Collections development – detecting and assessing where intentionality negotiates the boundary between mass (relatively indiscriminate) digitization and selective collections development
a. identify missing items, as when two of three volumes of a set are included
b. check against a few selected lists of desirable inclusions, e.g., APA fiche collection
c. note presence/absence of relevant foreign language materials
d. assess clarity and coherence of reported strategy of inclusion against what is actually presented by a collection

Potential Scholar-Participants
In light of the wide reaching implications of this project, we intend to approach scholars who have established names in more traditional research as well as those who have attained a national reputation for innovative work in the digital environment. Drawn from libraries and academic departments, an informal advisory committee will also be convened, with some teleconference calls and ongoing email exchanges through the various phases of the project.

Work Plan

1. Receipt of grant authorization
2. July 1-30:
    a. Finalize task description for scholar-investigators and review with a few volunteer colleagues from libraries and scholarly community
    b. Identify scholar-investigators for performing analysis
3. August 1- September 31:
    a. Scholar-investigators carry out the core work
4. October: Review of work product, draft report, some iteration of tasks to fill gaps and answer subsequent questions with scholar-investigators
5. November: meeting in DC hosted by CLIR with the scholar-investigators

6. December: preparation of final report, for publication by CLIR
7. 31 January 2008: report to Mellon.

Following the submission of the report, we would look to find appropriate venue and participants for meeting(s) to discuss the report among more senior scholars in various fields. Success of those meetings would guide selection of senior scholars to join an ongoing collegium.