The Core Historical Literature Approach for Selection for Digital Imaging

Joy Paulson

Since 1997, discussions about digital libraries have increasingly focused on the idea of mass digitization, that is, the creation of a critical mass of materials that is large enough to include all important publications in one or more subject areas. Google's December 2004 announcement of its intentions to digitize the collections of a number of major libraries is the most recent example of this approach.[1] However, in a paper presented at the Digital Library Federation Forum in November 2000, Carole Palmer, associate professor of Library and Information Science, University of Illinois at Urbana-Champaign, suggested shifting away from critical mass as a defining element and moving towards the goal of contextual mass, an approach that uses the essential features of scholarly work as its criteria for development.[2] Palmer wished to de-emphasize the size of digital collections, and place priority on the materials that scholars actually use.

This approach to selection for digital imaging is similar in many ways to the core historical literature approach to selection for preservation developed by Wallace Olsen in the early 1990s. Olsen, then a senior research associate at Albert R. Mann Library at Cornell University, pioneered the core literature approach in the seven volume series, *The Literature of the Agricultural Sciences*, which he edited.[3] He developed extensive bibliographies of scholarly monographs and serials for a specific subject and its sub-

[1] Google Checks Out Library Books, December 14, 2004, http://www.google.com/intl/en/press/pressrel/print_library.html (8 Mar. 2008).
[2] Carole L. Palmer, *Configuring Digital Research Collections Around Scholarly Work*, paper presented at The Digital Library Federation Forum, Chicago, November 2000.
[3] Wallace Olsen, *The Literature of the Agricultural Sciences*, Ithaca, NY: Cornell University Press, 1991 - 1996.

disciplines using citation analysis to determine levels of use over time. Extensive evaluative information was then collected from scholars, who provided their expert opinions regarding the relative value, historical importance, and potential future use of the historical publications listed in the bibliographies. This methodology resulted in priority-ranked lists, reflecting the understanding that, due to limited resources, it may not be possible to include all material published on a given subject in a digital library collection or to preserve all deteriorating publications. Initially, only those titles with the highest ranking were selected for inclusion in a digital library collection. Lower-ranked titles were included when there are sufficient resources to make this possible.

Olsen's approach was based on the principle that effective digitization and preservation plans should focus on a particular discipline as a whole rather than on the holdings of an individual library. No single library holds all the important publications in a specific subject area, regardless of the size or comprehensiveness of the collection. Since no single library is likely to have all the highly ranked titles in its collection, cooperative projects, having a number of participating libraries, are ideal. Although a single library can undertake this type of a project, it must be prepared to work with other libraries to obtain titles it does not own, which can be a time-consuming process, especially when attempting to assemble a continuous run of serials.

Mann Library has applied Olsen's core literature approach in creating two historical digital library collections: the Core Historical Literature of Agriculture (CHLA)[4] and the Home Economics Archive: Research, Tradition, and History

[4] Albert R. Mann Library, Cornell University, The Core Historical Literature of Agriculture (CHLA) http://chla.library.cornell.edu (8. Mar. 2008).

(HEARTH).[5] The United States Agriculture Information Network (USAIN), a community of land-grant universities, has also successfully used this method to identify and preserve the literature on agriculture and rural life for each state.[6]

This chapter will describe the process of developing a core bibliography, recruiting reviewers, reviewing and ranking the titles, and selecting materials for digitization from the ranked bibliographies. It will also discuss the strengths, weaknesses, and relative costs of this approach, particularly in relation to the mass digitization approach.

Developing a Core Bibliography

For the CHLA project, Mann Library adapted Olsen's core literature approach to historical agricultural literature. The most important pre-1950 scholarly agricultural literature was identified using citation analysis. The results of this analysis were used to create subject bibliographies of scholarly monographs and serials for each sub-discipline of agriculture: agricultural economics and rural sociology; agricultural engineering; soil science; crop improvement and protection; food science and human nutrition; animal science; and forestry. This method resulted in seven separate bibliographies, each containing a manageable number of titles for the scholarly reviewers to work with. For each bibliography, a group of scholars reviewed the titles and ranked them according to historical importance. Some titles were dropped on the recommendation of the reviewers, who also recommended titles be added to the bibliographies. The CHLA team

[5] Albert R. Mann Library, Cornell University, Home Economics Archive: Research, Tradition, and History (HEARTH) http://hearth.library.cornell.edu (8 Mar. 2008).
[6] *The National Preservation Program for Agricultural Literature*, 1993 http://preserve.nal.usda.gov:8300/npp/presplan.htm (8 Mar. 2008).

chose to digitize the most highly ranked titles in English published between 1850 and 1950, although some of the serials continued past 1950.

It is important to recognize that citation analysis is not always an effective approach for all disciplines. Citation analysis was not employed in the HEARTH project, since much of the historical home economics literature consists of practical works intended for a general—rather than scholarly—audience, and therefore often does not contain citations. This general literature is important for understanding the development of the discipline, and is of interest to scholars who specialize in women's studies and social history.

In the HEARTH project, the first step to creating project bibliographies for each of the sub-disciplines was to define the scope of the literature.[7] Specific subject parameters and criteria for inclusion and exclusion were established to ensure that the scope of each bibliography was clearly understood by those compiling and evaluating them. Because HEARTH was fundamentally designed for preservation, the project focused on materials produced between 1850 and 1950, a period in which cheap, highly acidic paper was produced, causing much of these materials to now be embrittled. The project also focused primarily on monographic and serial literature and excluded some materials, such as government documents, extension publications, dissertations, newspapers, manuscript and archival materials. The types of materials excluded from the project were materials likely to be included in other preservation projects, or materials of primary interest to a specific institution.

[7] HEARTH sub-disciplines: applied arts and design; child care, human development and family studies; clothing and textiles; food and nutrition; home management; housekeeping and etiquette; housing, furnishing, and home equipment; hygiene; institutional management; retail and consumer studies; and teaching and communication.

Titles were collected from home economics bibliographies; bibliographies in various subject areas, such as child development or nutrition; lists of works cited in recent scholarship on home economics, social history, and women's studies; and library catalog records. A database was created with EndNote software that allowed titles to be imported from online catalogs and bibliographic utilities. Monographic titles were identified using these sources and imported into the database from Cornell's online catalog or OCLC. The criteria for inclusion were applied as generously as possible, looking for any titles that may have been of importance to home economics, regardless of whether or not the authors considered themselves a part of this field of research. Multiple editions of titles were included, since there can be considerable variation between editions, and reviewers were allowed to determine which editions, if any, were of importance.

The HEARTH database contained over 13,000 titles. Each entry was assigned a code indicating the sub-discipline to which it belonged (some material was assigned to more than one sub-discipline), making it possible to generate a comprehensive bibliography for each subject area. These bibliographies varied greatly in length, from 200 titles to 2,800 titles. While the database contained only monographic titles, some serial titles were later included in the project. These serial titles were selected in a more informal manner. They were either generally recognized as important by scholars in the field or were specifically recommended by reviewers.

The USAIN Preservation Project has also used the core historical literature approach to identify, at the state and regional level, the most historically important literature on agriculture and rural life. As of 2004, twenty-seven states have participated

in this series of projects.[8] Each participating land-grant university develops a bibliography of the historical literature on agriculture and rural life for its state and region. Both serials and monographs published between 1820 and 1945 are included in the bibliography. There are strict guidelines for inclusion and exclusion in the bibliography. For example, no federal documents or manuscripts are included, since their preservation is undertaken by different sections of the National Preservation Plan for Agricultural Literature.[9] Materials published before 1820 are not included, since the National Agriculture Library (NAL) has taken responsibility for the preservation of this older material, while more recent material is not included, since it is not at this time as likely to be as embrittled as material published between 1820 and 1945. Extension materials are often excluded, since many states' extension materials had already been microfilmed in an earlier, NAL-sponsored project that ran from 1974 to 1987.

While the USAIN Preservation Project has thus far employed preservation microfilming as its reformatting technology, we expect that in Phase VI, which is due to begin in 2006, some of the participating libraries will choose to digitize their materials rather than microfilm them. However, we do not expect this decision to have an impact on the bibliographic phase of the project.

[8] The participating states include: Alabama, Arkansas, Arizona, California, Colorado, Connecticut, Georgia, Florida, Hawaii, Illinois, Iowa, Kansas, Maryland, Michigan, Minnesota, Montana, Nebraska, New Mexico, New York, North Carolina, North Dakota, Oklahoma, Ohio, Pennsylvania, Texas, Washington, and Wisconsin.
[9] NAL, *The National Preservation Program for Agricultural Literature*, 1993 http://preserve.nal.usda.gov:8300/npp/presplan.htm (8. Mar. 2008).

Hearth Case Study

Recruiting Reviewers

Before the bibliographies can be reviewed and ranked by scholars, the reviewing scholars need to be recruited. In selecting these reviewers, the needs of potential users should be considered first. CHLA and HEARTH are freely available on the Internet and have proved a valuable resource for scholars in various fields beyond those in agriculture and home economics, such as United States social history, the history of science, women's studies, and cultural studies. We also have a wide range of users from outside the higher education community.

One assumption of the HEARTH project was that a senior researcher in a specific field would be able to reliably choose the literature that his or her colleagues have considered most influential, but may overlook material that a historian might find valuable. Therefore, we recruited reviewers having a wide range of experience and subject expertise. For example, one of the sub-disciplines in the HEARTH project was "child care, human development, and family studies", and a number of scholars in this field, many of them retired or approaching retirement, were recruited for the project. Scholars in the later stages of their careers have had many years of experience; they understand how the field had developed over many decades; and they often have more time than their less experienced colleagues. Two historians were also invited to participate—one specializing in the history of science, and one in social history—since they would be able to offer different views of which titles were the most historically important.

The recruitment of reviewers was a significant and sometimes time-consuming task. We identified potential reviewers through contacts, such as advisory board

members, Cornell faculty, and librarian colleagues at other institutions. Recent scholarly literature by specialists within the sub-disciplines of home economics, and by historians, was reviewed, and the authors were contacted as potential reviewers. Bibliographers and collection development librarians with an interest in the historical literature of specific subject areas were often willing to serve as reviewers. For sub-disciplines that could be related to museum collections, such as clothing and textiles, curators were recruited. Scholars who had already agreed to serve as reviewers were often willing to recommend, or even actively recruit their colleagues.

Some projects using the core historical literature approach have been able to provide small honoraria, typically $400 to $500, for reviewers. However, this has not been true of all projects. The HEARTH project, for example, did not have any funds for honoraria for reviewers. It was simply necessary to rely on the willingness of busy scholars and professionals to volunteer their time. Fortunately, reviewers have usually been easily persuaded of the value and importance of the project, and we have not found lack of funding to be a major issue in recruiting reviewers.

Some reviewers have seen these projects as a means of increasing access to and preserving the historical contributions of their fields, a goal that seems to be especially significant for subject areas that are perceived to be marginalized within academia, such as home economics. Scholars in the humanities and social sciences sometimes see the prospect of easy access to deteriorating or rare source materials as valuable, both for research and teaching. In some cases, they have been eager to have access to a comprehensive bibliography in their area of specialization. Librarians and curators tend to be well aware of access and preservation issues, so it has not taken a great deal of persuasion to get them to participate.

While we typically recruited four reviewers for each bibliography, in some cases as many as a dozen reviewers were recruited. This was necessary for several reasons. Some sub-disciplines were very broad. It also seemed important to recruit non-academic specialists, such as museum curators to evaluate materials on textiles, or historians in order to ensure a broad historical perspective. This required the establishment and maintenance of relationships with a substantial number of people, which was a time-consuming but worthwhile endeavor. Later, when the collection was made available online, these same reviewers were able to help publicize and promote the use of the digitized materials.

The review and ranking process

Once the bibliography for a specific sub-discipline was completed—which was a somewhat subjective judgment—it was mailed to the reviewers in print form. A pre-paid return envelope was included in the package along with instructions for completing the review and ranking. Reviewers were asked to rank each title they recognized with a number from one to three, depending on their judgment of its historical significance. The following ranking scheme was used in this project:

1 = Very important historical title, of critical importance to preserve;

2 = Important title definitely worth preserving, funds permitting; or

3 = Worth preserving at some time, but of lower priority.

Reviewers were asked not to rank a title if they had no knowledge of that title and could not properly evaluate it. Specialization within a sub-discipline means that not all scholars in a particular sub-discipline are familiar with all of its aspects. We did this to try to avoid skew in the rankings if a scholar with no knowledge of a title gave it a low

ranking, while a scholar familiar with it gave it a high ranking. If none of the reviewers had any knowledge of a title, we did not consider the title as important for preservation.

When all the ranked bibliographies for a sub-discipline were returned, the results for each title were averaged, and titles were sorted and arranged into logical and manageable priority groups. For example, titles that fell into the 1.0 – 1.5 range became the first priority titles; titles that fell into the 1.6 – 2.0 range were ranked second priority for the project.

We have found it quite remarkable that the opinions of scholars within any specific sub-discipline were consistent. Among the multiple reviewers, the majority gave each title either the same rating or one rating apart, although there are some exceptions to this. We had initially expected a greater difference of opinion.

Selection of materials for digitization

Acquiring Titles

Titles with the highest rank became the first candidates for digitization. Project staff first identified highly ranked titles that were in Cornell's library collections. The Cornell University Library is made up of twenty different libraries, and materials were found to be in many different libraries throughout the library system. For the HEARTH project, titles were located in six different libraries, and in the library storage facility on campus.

In 2001, during the early stages of the HEARTH project, the scanning methods required that the volumes be disbound. Disbinding the volumes complicated the selection process and required project staff to work with the bibliographers for the various library collections to obtain permission to use the paper originals selected for the project. The importance of the project and the ranking of each title were explained to

each bibliographer, and information on the condition of the title was also provided. The bibliographers then decided whether a title could be included in the project. Most of the titles were printed on paper that had become embrittled, meaning that it would not be possible to rebind the volumes.

In 2002, our scanning vendor began to use an overhead digital camera and book cradle that met the resolution requirements of the project, and thus we no longer had to disbind each volume before scanning. This less destructive technology made it much easier for the bibliographers to make decisions about including material in the project. Some of the highly ranked titles were in Cornell's Rare and Manuscripts Collection and Mann's Special Collections. Decisions about including these titles were more difficult, and many of the titles were not included due to their age, fragility, and artifactual value. These titles may still be added at a later date, if they can be digitized in our in-house scanning facility, which would ensure their proper handling.

Cornell did not own a number of highly ranked titles, even though we have extensive holdings in these subject areas. For the CHLA project, titles not owned simply were not digitized, although they could be added at a future date when further funding is available. However, for the HEARTH project it was felt that the collection should not just reflect the holdings of a single institution. Two different strategies were undertaken to locate titles that Cornell did not own. During the period of time in which disbinding was necessary for scanning, we attempted to purchase titles through second-hand book dealers. This strategy was mildly successful, enabling us to include thirty-five titles in the project. However, many of the titles were unavailable, or else could not be purchased for a reasonable price. Once the transition was made to using an overhead digital camera for scanning, it became possible to borrow materials from other libraries. After

searching OCLC and RLIN, we requested permission from numerous libraries to borrow titles not owned by Cornell. The lists of titles needed were sent to the preservation or collection development librarians at the University of California, Berkeley; the University of Minnesota; the University of Nebraska; and Pennsylvania State University. Libraries at these universities collectively loaned several hundred titles to Mann Library. Other libraries indicated a willingness to participate in the project, but the project funds were expended before further titles could be borrowed.

The obvious difficulty with the core historical literature approach is that no library is likely to own all of the highly ranked titles identified in these comprehensive bibliographies. One option is to include lower-ranked titles owned by the library. Another option is to spend time identifying and negotiating with libraries that own the titles. Some of this work can be reduced by indicating, whenever possible, which library owns a title during the creation of the bibliographies. This is certainly possible for titles identified through the project library's catalog or OCLC and RLIN, but not for titles identified from bibliographies and citations. However, other libraries are often very willing to participate in projects of this nature, since the resulting digital collection will be freely available to their faculty and students.

Selection Impediments: Technology and Copyright

Technological considerations also play a role in determining which materials are actually digitized during a project. Difficulties are presented by the complex page layouts that are found in historical popular journals, which often have multiple columns, a variety of fonts, advertisements on the same page as articles, and articles that are continued in other sections of the issue. While it is possible to obtain good quality scans of the pages, Optical Character Recognition (OCR) software—which we have used to

enable searching across collections—does not handle complex page layouts well. These serials also tend to take much longer to structure and require many more judgment calls than most academic serials or monographs. Some material may be too fragile to withstand the digitizing process, or too rare to be digitized by an external vendor. Sometimes, these problems can prevent a title from being included in the collection.

Copyright plays a major role in the selection process, as well. Materials published after 1923 may still be under copyright protection. Including these titles in a digital collection requires a commitment to undertaking copyright review, and a commitment to working with publishers and other copyright owners to obtain permission to include the titles in the collection. Many digital library projects simply choose to avoid this complicated and time-consuming task by only including materials that are clearly in the public domain. However, the CHLA project team felt it was important to include all highly ranked titles, regardless of the copyright status. Therefore, copyright clearance played a significant role in this project. Sometimes the copyright owners were difficult to track down. Sometimes permission was denied. These obstacles, including the amount of time invested in this process, often determines whether a title can be included. However, many copyright owners of historical materials are eager to grant permission to include the title in a digital collection, although they often desire specific copyright statements and links to their websites.

Conclusion

While the core historical literature approach to selection for digitization is a time-consuming and labor-intensive approach to selection, and therefore relatively expensive, it can yield a collection of highly important historical materials in a specific subject area. At a time when libraries and their funding agencies have limited resources for digitization, but at the same time face increasing demands for online collections, this approach ensures that important historical materials are preserved and made more accessible. These smaller collections are easier to manage and preserve, and the total costs associated with these activities will also be lower than those for larger critical mass collections.

Focusing on a smaller collection of core titles may also make it easier to identify material useful to research, rather than sifting through huge amounts of perhaps less relevant material in critical mass collections. The input of scholars in the selection process means that both scholars and students view these core historical collections as highly desirable resources for their research, and ensures that these collections will be used.