Moving the Digital Library from "Project" to "Production"

John Price-Wilkin
Head, Digital Library Production Service
University of Michigan

ABSTRACT

The author explores the characteristics of a production digital library operation. He suggests that for digital libraries to mature, production organizations must be developed and that these production organizations must have certain characteristics. These production digital library operations must:

1. be fully integrated into their parent organizations;
2. have moved beyond a focus on "projects" to one of deploying and supporting classes of systems;
3. reflect a high level of investment consistent with the institutional mission;
4. undertake *long-term* maintenance and development of the collections and access systems they support.

Using as his example the University of Michigan Digital Library Production Service, Price-Wilkin examines production organizations and explores three areas of collection support that exemplify these characteristics.

INTRODUCTION

Digital libraries are in a period of exploration and rapid development. While our sense of what they are or should become varies, for the sake of this discussion we assume a definition consistent with that articulated by Don Waters, the Director of the Digital Library Federation:

> Digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities. [Waters, 1998]

While experimentation and research are essential to the development of effective digital libraries, it is ultimately formally defined and institutionally supported *organizations* that will ensure the viability of digital collections.

Wendy Lougee has suggested that we can apply a kind of human developmental model to the four distinct stages of digital library development, beginning with **Infancy** and culminating in **Maturity** [Lougee, 1998]. **Infancy** is marked by learning through projects. Projects not only facilitate learning, but test resistance to ideas, find opportunities, uncover resources and barriers, and eventually create stable building blocks. In **Adolescence**, we see peer modeling and the exploration of best practices: here, the digital library has begun to look outward both to other institutions, but also to methods and formats that will ensure longevity and interoperability. The third stage of development, which she calls the digital library as **Young Adult**, is where we find the organizations that I wish to explore in this paper. In this stage of development, we find an increasing focus on collaboration, especially for the establishment of standards and architectures for interoperability. **Maturity** is marked by the presence of a

fully functioning digital society, a market economy, and rich collaboration and knowledge environments such as those found in UARC [Finholt, 1995].

The relatively intense focus on digital libraries in the last few years has resulted in a handful of fairly mature production support organizations. Although experimentation for materials such as video continues to be important, these organizations have all successfully moved beyond the early stages marked by projects, work in isolation, and developing support for basic formats and methods. In fact, these mature digital library production organizations exhibit a new set of characteristics that are an important part of their productivity and livelihood.

1. They are not only engaged with the digital library community, but are fully integrated into their larger organizations (such as university libraries).
2. These relatively mature organizations have moved beyond a focus on "projects" to a mode where they are deploying and supporting classes of systems—systems into which more content can be added.
3. The production organizations reflect a high level of investment by the larger institutions in which they are situated. This investment is consistent with a recognition by the institution that the digital library is critical to its mission.
4. Finally, and significantly, these organizations undertake *open-ended* maintenance and development of the collections and access systems they support. While it is a certainty that the methods and strategies for maintenance will evolve, setting a term on the duration of responsibility for a collection can only contribute to a process of trivializing the collections of the digital library: these digital library collections are, in effect, our perpetual responsibility.

The University of Michigan's Digital Library Production Service is an operation that exhibits these characteristics. The discussion that follows is one that attempts to illustrate the value of these characteristics by exploring the organization of DLPS and three model systems that it uses to provide a high degree of functionality and cost-effectiveness in building the digital library.

BACKGROUND

The Digital Library Production Service is part of a larger University of Michigan Digital Library Initiative, led by one of the University Library's three Associate Directors. The UM Digital Library Initiative evolved from a set of successful but divergent campus digital library efforts in large part as the result of a 1991 Information Symposium, creating a campus commitment to building an "information agenda" [*Information and People*, 1991]. That Information Symposium, involving a broad cross-section of the University, culminated with a number of recommendations. The campus should attempt to:

- bring together library and technology expertise
- develop visible projects
- create an "Information Community"

Eventually, in support of these goals, the heads of the three principal information organizations on campus, the Information Technology Division, the School of Information, and the University Library, committed joint funding for a Digital Library Initiative. This funding included support for the creation of a position and funding for projects that would have campus impact.

Even this commitment was situated in a set of larger accomplishments, which allowed the formalization of goals for campus digital library initiatives. Prior to the Information Symposium, the University Library had begun to embark on some preliminary and perhaps formative efforts to build digital library components. Wide area access to non-bibliographic information sources—i.e., to the data themselves, rather than metadata—became a focal effort of the University Library in the late 1980's. The University of Michigan Library put in place a modest program of support for statistical data files, began exploring delivery of GIS data, and then ultimately put in place a formal access system for text encoded in SGML.[1] The Library had begun to build systems for storing and accessing library collections in electronic formats.

After the Information Symposium, beginning in 1992, the University Library undertook a number of initiatives that laid the groundwork for future work and future relationships. Some of these, like the Library's Gopher server, were unremarkable in their use of technology, but played important roles in information provision and partnerships. Others, like the UM implementation of TULIP, put in place significant pieces of infrastructure that would later prove instrumental in the University's development of digital libraries. The Library's Gopher server was remarkable in at least one respect, its aggressive approach to putting collection material online. Although it had previously undertaken similar efforts prior to Gopher (cf. UMLibText), the new mechanism allowed Library staff to mount collections like statistical information from the Commerce Department in ways that reached extremely broad audiences [York, 1994]. The strong presence of the Library's Gopher on campus also contributed to early cooperative efforts with the University's computing organization, ITD. TULIP's contributions were even more profound, if less visible.[2] Although Michigan's was one of several Elsevier TULIP implementations, at the University of Michigan TULIP spawned a development effort that saw the creation of FTL, the search engine now used by JSTOR, and tif2gif, Doug Orr's optimized GIF generator for TIFF G4 images, now used by the University of Michigan Making of America system.

A broad, campus-wide partnership has played a powerful role in moving the concept of digital libraries forward at the University of Michigan. As mentioned earlier, the Library's Gopher project played an important role in establishing a partnership between the University of Michigan's Library and the Information Technology Division. Like Gopher, the TULIP effort contributed to bringing the Library together with another campus entity, the School of Information. These partnerships were formalized in 1993, and Wendy Lougee was appointed the director of campus-wide digital library efforts. That partnership is a powerful presence today in the University of Michigan's digital library work, with management and advisory committees composed of representatives from each of the organizations, and multi-organizational funding for initiatives.

A number of significant UM digital library efforts began in 1994. Among these is the now organizationally independent JSTOR, which grew rapidly in size and scope, and now plays an important role in shaping our expectations about archiving and retrospective conversion [Guthrie, 1997]. Another major initiative appearing in 1994 was the UM Humanities Text Initiative, which served to expand the earlier UMLibText effort and provide a significant

---

[1] Numeric data were made publicly available via FTP (see York, 1998), and user support for codebook use and modest data extraction were provided. The UMLibText is described several places, including Price-Wilkin, 1991.

[2] For a summary of the TULIP project, see *Library Hi-Tech* 13:4 (1995), which includes a series of articles on the topic.

WWW presence for the University's SGML-based text collections [Powell and Kerr, 1997]. Also appearing in 1994 was the University of Michigan's NSF/NASA/ARPA-funded digital library effort, focusing not on production technologies, but on the role of agent technologies and distribution of responsibility for the digital library [Durfee, Kiskis and Birmingham, 1997]. The breadth of these 1994 initiatives provides some indication of the fruitfulness and variety of the evolving environment at Michigan. (figure 1)

It was this growing proliferation of significant digital library activities—activities that had moved well beyond the "experimental" and had begun to reach wide audiences—that contributed most to the recognition of a need for a digital library production organization. Between 1994 and 1996, while the activities mentioned above continued to grow, new efforts at Michigan were introduced, and along with them new models and formats. Among the significant efforts undertaken were UM's MESL implementation, the UM Making of America development, and negotiation of the Elsevier journal content that would later go into the PEAK system [Stephenson and McClung, 1998; Bonn, 1999; Mackie-Mason, 1997]. These initiatives will be discussed in greater detail later, but they brought to the digital library environment at Michigan a variety of significant new elements. Not only were new formats or methods introduced (continuous tone images in the case of MESL, and preservation-oriented monograph conversion in the case of MOA), but several of these efforts helped to highlight the absence of a formally defined organization to support the effort. In 1996, the group that guided (?) Wendy Lougee in campus digital library efforts embraced a plan for, and then committed the resources needed to create, a digital library production service.

FORMATION OF DLPS

ORIGINS

The University of Michigan Digital Library Production Service was formed in 1996 through a commitment of resources from the four organizations then guiding the campus Digital Library Initiative. The Information Technology Division allocated a substantial sum from its personnel budget (and many early DLPS staff were hired from ITD). The newly formed Media Union, home to the University's advanced and robust CAEN computing organization, provided DLPS with hardware and software, staff support for hardware, and use of a climate-controlled facility with around-the-clock support. The School of Information and the University Library provided a combination of staffing, hardware, and space (as well as funds for renovation of the space). Additionally, the University Library fully funded all digital collections and most of resources necessary for creation of new materials. Formed from resources from the four organizations, DLPS works with each to ensure that it is meeting the goals set out by each. For example, DLPS supports major collections used by faculty and students in Schools and Colleges programmatically linked to the Media Union.

DLPS was established with clearly defined areas of responsibility. Not only would it provide long-term support to the growing array of production digital library operations (e.g., the Humanities Text Initiative) at the University of Michigan, but it would undertake a process of articulating and implementing a number of higher level goals. DLPS was made responsible for defining near-term digital library architectures for the campus, primarily refining those mechanisms it had already put in place, and extending them to create a more fully integrated environment. Similarly, it would work to take lessons learned in previous efforts to define appropriate document or data structures for the digital library. This goal was seen as essential for ensuring that our investments made in digitization would have enduring value. DLPS was also made responsible for application development and maintenance in those primary areas of

responsibility (e.g., bitonal page image systems, continuous tone color image systems, and encoded text systems). As new formats such as video evolve, it is expected that DLPS will take responsibility for them as well. Finally, DLPS was charged with responsibility for basic operations such as data loading, and ensuring that servers and software have appropriate levels of maintenance.

INTEGRATION INTO THE LIBRARY AND RELATED ORGANIZATIONS

The University of Michigan Digital Library Production Service was *not* intended to operate self-sufficiently, in isolation of the contributing organizations. There are important ties to each of the other organizations, and especially the University Library. For example, DLPS relies on the Information Technology Division and the Media Union to provide a high level of computing support, which those organizations view as critical and enabling infrastructure. The services are provided in a manner consistent with the mission of the digital library (e.g., uninterrupted support for servers, as well as unique security and operating system configurations), ensuring that DLPS staff are devoted to building and maintaining digital library collections, rather than the hardware on which they reside.

Mainstream Library staff members provide an array of services critical to the operation of the digital library. Public service staff members provide user support for online collections, as well as end-user instruction. Collection development staff members are responsible for selecting digital collections for local deployment and work with DLPS staff on "E-Teams" to weigh the advantages and disadvantages of alternative means of delivery. Preservation staff, with guidance from DLPS, make determinations of the most appropriate means of digital capture, and then prepare the materials for digital capture (occasionally operating equipment for the actual capture). Mainstream Cataloging staff members create descriptive metadata for locally-converted materials, and specialized digital metadata specialists in the Cataloging Department help guide DLPS in decision-making for mapping between standards, display of bibliographic elements, and related issues. Similarly, DLPS maintains important relationships with other areas of the Library, including the Library Systems Office, Acquisitions, and Special Collections. Although DLPS staff members have responsibilities that touch on all of these areas of library operation, the intention of this design is to ensure that the most qualified staff member performs each task; the intention is *not* to recreate the Library within DLPS. Consequently, DLPS is fully integrated into the entire Library operation.

MISSION

Like its staffing, the goals of the Digital Library Production Service are more specific than the larger organizations within which it exists, including those of the campus-wide Digital Library Initiative. The campus-wide DLI has responsibility for creating a collaborative organization and a coherent environment for networked information, whether locally-maintained or held remotely. Further, the DLI has a broader mission in its responsibility to explore, for example, economic and policy issues for the campus and beyond. By contrast DLPS' responsibilities are constrained to information that is held locally (either at the University of Michigan, or at other institutions where we are assisting in implementations). It is responsible for **designing, creating, and maintaining** the mechanisms needed to deliver

library information via networked mechanisms.[3] Where possible, DLPS works to provide coherence to distributed activities on campus, typically by federating collections "owned" by organizations outside the Library. It is important to note that DLPS respects the autonomy of those collections, offering incentives to join in federation and ensuring (where possible) that those collections can be accessed independently through mechanisms supported by DLPS. Also part of this larger goal is DLPS' responsibility to articulate an "information architecture" for the campus in ways that position the University to participate in inter-institutional cooperative activities. To this end, for example, DLPS has been active in a number of national and international digital library initiatives.

STAFFING

Staffing in DLPS has grown with its accomplishments. Initial staffing was set at levels necessary to provide a baseline of commitment to all of these areas, with growth expected for new formats and for extending DLPS' commitment to issues such as cross-collection/format integration. In 1998, the University's Provost made arguments on behalf of the Library's budget request, allowing both a consolidation and an extension of funding for DLPS. Currently, DLPS is approximately twenty full-time equivalent (FTE) staff (with only two FTE in partial and student appointments).[4] As of early 1999, there are four primary areas of DLPS (figure 2):

- Two areas within DLPS are *format-specific*, with responsibility for complex and specialized decisions surrounding the selection of formats, their application, related online implementation issues, and long-term support for these formats. Currently, these two areas are encoded text (SGML and XML) and continuous tone imaging. Chris Powell, the coordinator of the Humanities Text Initiative, has responsibility for the first of these areas, and has staff and resources to ensure that texts are created in appropriate ways and that collections are mounted effectively. John Weise, the coordinator of Image Services, has responsibility for continuous tone imaging, and similarly has staff and resources for creation and conversion, as well as online systems.
- The third area within DLPS is a relatively large and well-integrated Systems Group. The programmers within this group work in close cooperation with the service coordinators (above) to ensure that systems are created to effectively support online access to collections in these and other areas. All systems are built through a team effort, drawing on staff from across the organization, and especially those within the Systems Group. Many programmers within the Systems Group have areas of specialization, including (for example) SGML and XML. This ensures a high degree of technical *and* format understanding in building the online systems.
- The fourth area within DLPS is an "infrastructure" group, including staff members who provide services that touch on most areas of DLPS operations. For example, a DLPS Interface Specialist is responsible for working with implementation teams to perform a functional "needs assessment" for each system, and then to design and evaluate the

---

[3] This responsibility excludes information that is simply bibliographic. The University Library's Systems Office maintains a variety of systems for creating, maintaining, and accessing bibliographic information, including the Library online catalog, MIRLYN.

[4] An important part of DLPS' growth in 1998 was the establishment of a development organization. Also a part of the Library's larger Digital Library operation, Digital Library Program Development has as its responsibility the development and testing of models and programs that may eventually find their way to the production organization. This expansion of the UM Digital Library Initiative will be extremely helpful in allowing DLPS to focus its energies. While DLPD has begun as a relatively small organization (approximately four FTE, including one FTE of student support), we hope that it too will grow as opportunities and challenges present themselves.

interface for that system. Also in this group are staff members who perform data loading and technical support for DLPS staff. This group is led by an assistant head of DLPS, who also has as a major responsibility for DLPS "data" management. This critical area provides an integrative oversight role for data creation and management, giving attention to questions such as what data elements are captured, where are they stored, and how can we ensure effective migration of data.

A fifth area of DLPS is the NEH-funded Middle English Compendium operation. The MEC staff members have responsibility for converting and encoding the monumental *Middle English Dictionary*, and creating the complex and authoritative HyperBibliography of Middle English Prose and Verse. The MEC, online at http://www.hti.umich.edu/mec/ represents a major effort in Middle English philology, and intends one day to add a large body of encoded text primary sources from the period. Although the future of the MEC is still uncertain, we are hopeful that it will generate sufficient revenue to support the ongoing extension of the resources it includes.

COLLECTIONS

Currently, DLPS manages collections in three principal areas: bitonal page images (typically with raw text to facilitate retrieval), encoded text with or without associated bitonal page images, and continuous tone image collections. Within each of these areas are book and journal collections, and a variety of art, architecture, and artifact images. Including the pages in the PEAK collection (Elsevier journals), DLPS manages approximately eight million pages of books, journals, and reference materials online. In 1999-2000, we will add to this approximately 2.3 million page images as part of a monograph conversion project using titles in our Buhr Storage facility, as well as several million more pages of journal material. Particularly under the aegis of the Humanities Text Initiative, approximately three million pages managed by DLPS are encoded text. Most of this material is historical literature (fiction, poetry, philosophy, and other prose), but a significant portion of the SGML/XML material is also reference works such as the *Middle English Dictionary*. Image Services manages a diverse body of materials, ranging from high resolution images of papyri from the University of Michigan Papyrology collection, rare books and manuscripts, archeological objects from the Kelsey Museum of Archeology, and art and architecture images from several campus and off-campus organizations. In all, it contains approximately 25,000 images, and 100,000 descriptive records for visual resource materials. It is important to note that though these resources are nearly without comparison in the digital library community, the size of these collections—approximately 1,200 journals and 20,000 monographic titles[5]—they pale in comparison to the large research library.[6]

ECONOMIES OF SCALE

The size and array of resources within the University of Michigan Digital Library Production Service creates significant economies of scale, making many processes less expensive and ensuring the availability of resources to systems or activities that might not otherwise have them. We are able to capitalize on investments such as expensive search engines (or "free" search engines that are expensive to support), using them for a variety of projects ranging considerably in size. Some examples of these sorts of economies are:

---

[5] The size of the visual image collections is growing quickly and may soon be a size comparable with campus slide collections.

[6] Current collection statistics, including coarse indicators of usage, are online at http://www.umdl.umich.edu/dlps/stats/.

- **"Persistent" URLs**: The Internet still lacks a dominant and successful model for creating URNs, leaving digital library production organizations with a need to invent and sustain interim methods. At DLPS, these methods (based on CGI-accessible database mechanisms) are elaborated once for each type of system supported by DLPS, and are then used subsequently as collections are added.
- **Internet commerce and authentication**: Investments in methods for authentication and commerce are ultimately necessary for organizations of this sort. At DLPS, we have used several methods for per-item commerce transactions, including credit card transactions, and use Oracle databases for multi-institution authentication in the PEAK system. These mechanisms, along with SSL, are available as part of other systems or collections that DLPS builds.
- **Tools and methods**: Like the resources needed for authentication, many specialized and often expensive tools are needed for digital library support. Search engines and compression/decompression tools for on-the-fly conversion are just two examples. DLPS uses tools such as tif2gif, CPC, and wavelet compression are used in a variety of systems as dictated by need and format.[7]
- **Expertise in complex areas**: Skills in challenging areas such as SGML/XML or support for SQL databases are an important commodity, and building a production organization by "leasing" such expertise is not realistic when the systems of access or management depend so fundamentally on them. DLPS has benefited from these sorts of investments, especially in SGML/XML skills and in imaging (formats, color, and metadata).

DLPS ARCHETYPES: SYSTEMS, *NOT* PROJECTS

The University of Michigan Digital Library Production Service is working to define classes of systems, and to put in place mechanisms to support these classes. These systems represent archetypes, and each collection is assessed to determine whether and how it might fit existing systems. We are careful not to *force* a particular body of material to fit one of the existing systems, but when it is possible, we are able to achieve those previously mentioned economies of scale much more effectively. This approach also results in other significant advantages.

- By approaching support for the digital library in this way, we are able to draw on significant format expertise within DLPS.
- We are able to deploy collections in an environment of ongoing support and planning. Each class represents a model or framework for continuing expansion rather than a standalone system that could be orphaned.
- DLPS staff members are able to participate more effectively in national or regional cooperative efforts, sharing models and expertise. These larger efforts enrich DLPS efforts at the same time that DLPS contributes to the larger digital library community.

---

[7] Tif2gif is a locally-developed tool for very fast conversion of TIFF G4 images to various levels of GIF. Written by Doug Orr, it is optimized for the unique characteristics of TIFF G4, and generates GIF images so quickly that it is not necessary to store pre-computed GIFs. Source and binaries are available online at http://polyphemus.engin.umich.edu/tif2gif/. CPC is Cartesian Inc.'s compression format for TIFF G4 images. Using it for the PEAK system, for example, we are able to achieve a nearly 3:1 compression ratio. Conversion from CPC is sufficiently fast that we are able to convert images to other (user-requested) formats on the fly. Information on CPC can be found at http://cartesianinc.com/. To provide greater functionality for continuous tone images, we use LizardTech's Mr. Sid wavelet compression software. Using wavelet compression, we are able to store the full resolution of the image, albeit compressed, and from it we can generate derivative resolutions or enable panning and zooming on portions of the image, all in real time. Information on Mr. Sid can be found at http://www.lizardtech.com/.

As we continue development of these classes, we are frequently able to see opportunities for consolidation or expansion. A powerful example of this refinement process was encountered in 1998. The HTI, in cooperation with SGML Systems, deployed each new collection in its native SGML encoding, re-using access mechanisms previously developed wherever possible. Although DLPS principles of re-use were in wide evidence, the large number of legacy systems for text collections had become a significant problem when we wanted to add new features. Each new feature assumed a foundation of code that may or may not be in existence, depending on when the collection was put online. A strategy was articulated for encoding all such collections in a common format, and building a single piece of code to provide all of the functionality needed for all of the text collections. If the collection was predominantly drama and HTI wished to provide searching on speakers and speeches, this subroutine would be invoked; otherwise, in cases where it was inappropriate to the collection (or speaker and speech were not encoded), speaker/speech searching would not be made available. The approach was christened GUMS (General Unified Markup System), and early work on moving to implement collections and systems using GUMS has been very successful. There is a clear sense, however, that we must undertake an ongoing activity, parallel to the work of extending the current systems, to refine and document of the current set of classes.

THREE SAMPLE MODELS

The principles elaborated earlier are clearly visible in many of the collections or services within DLPS. Each of the following three examples benefited from previous work within DLPS, and each forms a class, as well as a system to which more content can be added.

Making of America and Encoded Text Systems

The system built for the first University of Michigan Making of America collection was designed to fully support preservation-oriented conversion methods and provide a gateway or even bridge to the high-quality access systems developed for the Humanities Text Initiative. The history of the UM MOA development effort is documented elsewhere, and a more lengthy treatment of the technology used is also available.[8] The UM MOA system has been extraordinarily successful, both as an architecture within the DLPS digital library efforts, and as a digital library resource for a broad range of users. (figures 3-6)

Of primary importance in the UM MOA architecture is support for the products of a Library Preservation process. Images in the system are 600 dpi bitonal, TIFF G4 files. No derivatives (e.g., GIF or JPEG images) are created or stored, except at the time of viewing request. When a user requests a page, the system generates a GIF or PDF derivative in real time and without any appreciable delay (typically less than one second). Four levels of resolution in GIF are made available to users, taking into account the wide range of displays and network connections; a 600dpi PDF version is also made available, primarily for printing. (figure 6) While the number of pages—approximately 3 million by late 2000—is relatively small compared to a typical research library collection, its large size, expected continued growth, and continuing changes in desktop technology (including networking) argue against storing anything but the master images online. Use patterns also suggest that as long as we are able to generate appropriate derivatives in real time, based on user demand, we will significantly minimize the requirements for management [Price-Wilkin, 1997]. Of course there are still concerns about the appropriateness of TIFF G4 as a preservation-quality

[8] For the history of the UM MOA effort, please see Bonn, 1999. For information on the technology used in UM MOA1, please see Shaw and Blumsen, 1997.

surrogate for pages, but the University of Michigan Library believes that this format provides a high quality surrogate for most printed materials.[9]

Page images from the Preservation conversion process are subsequently treated by automated Optical Character Recognition (OCR), and the OCR is associated with the page image using a simple form of SGML. The extensibility of the MOA system has in fact been tested through our OCR processes: two generations of OCR have been applied to all of the materials without the need to change the system architecture. Current OCR technology used by DLPS is a voting system, providing the MOA system with a significantly high quality of character representation than typical OCR.[10] The system exhibits approximately 99.8% accuracy for nearly all content-bearing pages (e.g., excluding pages with engravings and textual pages such as the title page and advertisements) [Bicknese, 1998]. The SGML applied to the text is XML-compliant, and provides information such as image location, page type (e.g., table of contents), "confidence" of the OCR, and page number. (figure 7)

The automatically generated SGML in MOA is also largely consistent with the Text Encoding Initiative (TEI) Guidelines, allowing a full integration with DLPS' encoded text efforts. While certain attributes such as those mentioned above (e.g., OCR "confidence") have been added to the MOA SGML Document Type Definition (DTD), it is otherwise entirely consistent with the TEI. This has allowed DLPS' encoded text operations in HTI to extract individual texts and upgrade them, correcting OCR and applying fuller encoding. Because the MOA system is fundamentally TEI-compliant, it can accommodate both the loosely encoded texts and texts with more detailed encoding. As resources are available to HTI, materials can be enhanced in MOA, ensuring better retrieval and higher levels of functionality for users. (figures 8-9).

Extensibility is critical for a system of this size and importance. The UM MOA system has been designed to be augmented in a variety of ways without significant overhaul. For example, DLPS has regenerated OCR for MOA without interruptions in service, thereby improving retrieval. Texts can be augmented by HTI, as discussed above. New texts can be added as Preservation resources allow, and all indexing and preparation of Web pages is automated. The underlying body of materials can remain largely unchanged while work on enhanced interfaces takes place—a process that took place between November 1998 and January 1999. We believe that the MOA system is a model of scalability and extensibility.

The success of the MOA archetype has been multifold. It has allowed us to add materials, a process now underway with MOA4 (2.3 million additional pages). It has helped reduce HTI costs, providing a readily accessible surrogate for the encoding process; in 1998-1999, HTI will add 100 more texts to the American Verse Project collection (see http://www.hti.umich.edu/english/amverse/). HTI works with the MOA system and contract services to keep its overall cost down and ensure future integration of its products with a Preservation surrogate. Perhaps most significant among MOA's successes, however, is the level of use by a broad range of users. While the printed source materials were largely unused (most had not circulated in more than ten years), in their online format they are

[9] The author would like to suggest that this format is highly successful as a surrogate for printed materials, and that energies in the Preservation community should be focused on a plan to coordinate storage of the books and journals themselves, *in conjunction with* this method of digital conversion. No method of surrogacy can *replace* the printed artifact, but the method used for projects such as MOA will cost-effectively satisfy nearly all user needs, leaving the original to satisfy the remainder.

[10] We use Prime Recognition's PrimceOCR. Information is available at http://www.primerecognition.com/.

searched some 100,000 times each month, and approximately 100,000 page images are displayed. The constant stream of positive user responses comes from genealogists, philologists, and academics alike.[11]

Image Services

The system built by DLPS Image Services provides a high degree of functionality for a variety of images of objects ranging from art and architecture, to artifacts, to manuscripts. It is also a model for federating diverse collections, respecting the autonomy and heterogeneity of the source collection while creating a union collection for mainstream users.

Image Services uses a mapping strategy for its metadata, keeping the native field names and data but simultaneously "tagging" each field with a corresponding value in a sort of Dublin Core scheme. This allows Image Services to provide collections in two ways: first, as a highly functional version of the database maintained by the contributing organization (e.g., the University of Michigan Museum of Art); and second, with generic field labels common to all of the images in the system. Thus, for example, the Museum of Art might use the field label "Artist," while the History of Art Department might use the field label "Source." Having selected only the Museum of Art data online, the user would be presented with search options for "Artist," and data would be displayed accordingly. However, if the user search across the entire collection, the Dublin Core value "Creator" is presented to the user as a search option and in the display of records. (figure 10) This, along with features for tailoring display and functionality (discussed below) ensures that a contributing organization such as the Museum of Art will enjoy the benefits of a powerful and highly functional system for their data. Trying to bring together a wide variety of administratively separate collections on campus is challenging, but we believe that the incentive of a cost-effective, inexpensive, highly functional host service (rather than, for example, administrative mandates) has been very effective in creating a unified database managed by DLPS.

DLPS Image Services uses a unified database to ensure a high level of performance and flexibility. Indeed, many of the collection providers maintain local management systems that could be brought online successfully. The diversity of systems, including Oracle, FileMaker, Embark, and others, would present a serious challenge to creating a distributed search feature. Instead, data are extracted from each of these systems on a periodic basis, and are then ported into the DLPS Image Services system. (figure 11). Using the methods described above, Image Services is still able to provide access to each collection as if it were an isolated, separately managed collection; the federating approach does not force a lowest common denominator effect. Instead, the significant resources of DLPS can be leveraged to provide faster search mechanisms, low-cost RAID, and around-the-clock maintenance of the database. These too are effective incentives to organizations when considering whether to mount their data through a central campus agency such as DLPS.

---

[11] Among the many positive user responses that continually come to the MOA site, this remark by the chair of the History Department at a California university is one that staff members have most appreciated:

> I have just briefly tried MOA and it is the most amazing, spectacular research tool since the Xerox machine. It is what I assumed the future of libraries would be, but to be quite honest, I never believed I would live to see so much of the past put on-line in such an accessible form. Business data sure, but history?? 'Paradigm shift' is almost too limited a term. To be able to search for any word and pull the document up on the screen (and print it out) boggles the mind. I have a book at the publishers now, and realize that I am going to have to pull the manuscript until I get a chance to use your database. Congratulations, your founders have the thanks of professional historians and students for their foresight into what is clearly the future of the past.

The Image Services system is very rich in functionality. Primarily to satisfy the variety of needs created by the diverse collections, Image Services has built its system using a template system. By specifying (and possibly even creating) a different template, one can make the system appear radically differently. For example, the primary interface presented to users is a search interface, with results appearing as a collection of thumbnails with associated descriptive information. (figure 12) Each thumbnail and descriptive label is, of course, a link to a fuller resolution view and more descriptive data. However, by invoking another template such as that for "slide shows," the system can be made to appear as a set of larger images in a pre-selected and set sequence. (figure 13) Another templates provides an interfaces for comparison of multiple images, critical for use of art and architectural images. (figure 14) Each high-resolution image is ultimately displayed with a pan-and-zoom interface. (figure15) Users can take advantage of low network speed connections or low-end displays by sending relatively small segments of high resolution images across the network, and panning directionally to view more of the image. This feature has been extremely useful to a wide variety of users, including our Papyrologists, who use the system to work with extremely high resolution images of papyri, and are rarely equipped with computing resources to be able to bring up the full resolution image.

The approach taken by Image Services is extremely scalable. Adding other collections incurs only small, marginal costs (e.g., RAID and time to process the new collection) rather than requiring us to build a new system for each. We are also able to add new functionality easily, adding "modules" or subroutines to the current middleware rather than re-writing the programs each time. At the same time, the system provides a high degree of functionality for a wide variety of users, including those who own and maintain the collections represented.

PEAK[12]

The PEAK system is, in the simplest terms, a mechanism for delivering the entire body of Elsevier Science's journals to a dozen institutions, with varying levels of access within and across institutions. It is, however, an excellent example of the value of a production organization in an academic research institution. The University of Michigan was able to use DLPS to leverage production technologies in support of the research mission of the University. Although alternatives were available to using a locally-developed system, those alternatives would not have allowed us to use the Elsevier materials to learn about the economic decisions of users and institutions when using current journals.[13]

Critical to our being able to support the research mission was our ability to use known technologies to bring the journals online. The University of Michigan had prior experience with the Effect Specifications (i.e., Elsevier's format for delivering image, OCR, and metadata) during the TULIP experiment, and as mentioned earlier, UM developed several significant tools to make the TULIP journals available. Notably, Ken Alexander had developed the search engine, FTL, which was subsequently used and refined in JSTOR, and Doug Orr had developed tif2gif to enable real-time generation of GIF derivatives from the TIFF G4 images. (figures 16-19) The use of these tools in previous environments (e.g.,

[12] This is intended primarily as a description of the way that the production technologies aided in pursuing a research agenda. A fuller report on the research conducted in PEAK should be available in mid-1999.

[13] Moreover, the commercially available systems lacked important functionality. For example, in developing our own system we were able to easily integrate CPC compression to save disk space, and UM's is the only working system (offering Elsevier's contents) that indexes entire issues, including miscellaneous front and back matter such as indexes.

TULIP and the early days of JSTOR development at UM) allowed us to look past these critical hurdles and instead to focus on supporting the research model. The availability of the known technologies also allowed us to invest energies in putting in place database mechanisms for authentication and for subscription/purchase control information, as well as new methods for compression to handle the large amounts of material. Material began to arrive in late summer, 1997. Within a few months, the system was ready for use at the University of Michigan, and by the beginning of 1998, it was released to the subscribing institutions.

With Elsevier's cooperation, Michigan contracted with eleven other institutions to deliver the 1,200 journals with some very unique subscription models. Using a model dictated by Prof. Jeff Mackie-Mason, four types of access were put in place:

1. **Free:** Older materials (i.e., pre-1997) were made freely available to all project participants. All project participants were also permitted free access to bibliographic and full-text searches, with no charges levied for viewing citations or abstracts.
2. **Traditional subscription:** As its name suggests, this model corresponds directly to the familiar subscription by journal title. A traditional subscription to a journal title permits unlimited access to the articles within that journal for the subscriber or subscribing institution.
3. **Per-article purchase:** When access to an article is purchased by a user, the user has unlimited access to that article. While at one level, this is a fairly unremarkable feature, Mackie-Mason's additional stipulation that this access should not be time-bound added the burden of using a database to associate a user identity with an article identity. This ensured that that the user who purchased access to an article would continue to have unrestricted access to the online article for the duration of the project.
4. **Generalized subscription:** The generalized subscription is the most novel and most challenging addition to PEAK. A generalized subscription consists of a number of "tokens" that can be spent on articles. An institution would purchase "bundles" of tokens, or generalized subscriptions. When tokens are available, they are spent on behalf of users for articles not covered by other subscription models. For example, if a user attempted to read an article from a journal not in an institution's traditional subscriptions, a generalized subscription token would be spent on that article. The article would then be available, without restrictions, to all other users at the institution. Again, the PEAK system must collect information about individual article access permissions, and associate them with users from each institution.

The research model further complicates these methods for access, where all methods for access are not available to all institutions, and not all institutions choose to take advantage of all methods available to them. This creates a complex matrix of users and materials, a matrix that must be available and reliable for the system to function properly. (figure 20). Please note that publications analyzing the results of this research will follow. We have already learned important lessons, however, and are working to share data on use with the participating institutions.

In implementing PEAK, our production technologies *and especially our production organization* allowed us to extend the digital library more fully into the University's mission of research and teaching. Independence from Elsevier was critical in order for us to be able to test these models, and the body of Elsevier materials was equally important to ensure that users would have a valuable body of materials that would draw them into the research environment. The ultimate control and flexibility of the local production environment allowed the University of Michigan to perform research that would probably not have

otherwise been possible, or could not have been performed in ways that the researcher stipulated.

MODELS—CONCLUSION

These model systems help us to understand the value of a production organization in building the digital library. They also amply demonstrate each of the principles or ideals proposed as critical for the functioning of a digital library production organization.

1. None of these systems would be possible without the full integration of DLPS into the Library, and indeed the University itself. Politically and procedurally, the fact that DLPS is a part of the University Library ensures a range of outcomes, including effective preservation-oriented conversion and deployment (in the case of Making of America), online strategies guided by information professionals (in all of the systems), and integration of online resources with subscription and acquisition procedures (especially in the case of PEAK).
2. Each of these systems is or represents a "class" of information, with consistent characteristics and support. Each is easily extensible through the addition of features or migration of code, and each continues to grow (in size) without revision of the underlying mechanisms. The imperatives of DLPS have ensured that we have invested in these models and long-term value, rather than hastily deploying "stovepipes" of separate and inconsistent systems without integration.
3. The investments made in the systems (and support for them) reflect a high level of investment consistent with the institutional mission. The quality of the underlying resources, for example, is extraordinarily high, just as the underlying metadata are extraordinarily rich. The staff resources (20 FTE) are primarily base budget resources, and staff skills are very high.[14]
4. All of the systems were constructed so that they could be supported in an open-ended fashion (i.e., as much as it is possible to say this, in perpetuity). Capture formats are all standards-based and high fidelity; in fact, most are suitable for creating replacement copies of original publications. In nearly all cases, the "archival" version of the digital surrogate is also the online version.

These models are all cost-effective and sustainable, and represent model systems for DLPS, with a significant emphasis placed on long-term support and extension.

CONCLUSION

The progress of digital libraries will depend increasingly on the creation of production organizations, particularly within libraries. The digital library cannot thrive on a proliferation of experimental projects, projects that will tent to undermine the viability of collections and work against cooperative architectures. In order for the digital library to thrive, we must begin to see a shift from "projects" to "archetypes." A focus on archetypes rather than projects is only possible in a production organization because of the need for continuity and re-use of resources in a coordinated, planned way.

The effective digital library production organization must be fully integrated into campus' academic mission, and especially into the mission and functions of the library. This can only be done by situating the digital library production organization *in* the library. Too many of the processes and resources needed to support the digital library are a part of libraries; the principles for information organization and management are an essential part of librarianship.

---

[14] While not necessarily an indication of skill, it is useful to note that most DLPS staff members are at relatively high levels of classification.

Success in creating digital library production organizations like those I describe here will also lead to an increased probability that we will successfully federate digital library resources. The holistic approach creates not only economies of scale, but also important opportunities for integration. At Michigan, this approach is leading to the elaboration of an architecture where every digital object is managed in highly functional ways that ensure the long-term maintenance of that object. This architecture also brings these resources together in ways that are transparent to the user, but which ensures tight integration of multimedia resources (figure 21). A focus on overall architecture is essential, and again can only ensue from testing hypotheses in production organizations.

Finally, sustainability is one of the key issues of the digital library, and an issue that also argues for the presence of the production operation. Certainly, as an ideal, many of us can readily embrace the notion that decisions must be made with regard for long-term value of the digital object. It is only through a permanent production organization—an organization with funding in a base budget, with open-ended appointments for staff, and with long-term responsibility for maintenance and migration—that this ideal can be supported.

BIBLIOGRAPHY

Bicknese, Douglas. "Measuring the Accuracy of the OCR in the Making of America: A report prepared by Douglas A. Bicknese in fulfillment of Directed Field Experience requirements, Winter 1998, University of Michigan, School of Information. See http://www.umdl.umich.edu/moa/moaocr.html.

Bonn, Maria. "Building a Digital Library: The Stories of the Making of America" Forthcoming in *The Evolving Virtual Library: More Visions and Case Studies*, Laverna Saunders, ed. Information Today, Inc. See also http://www.umdl.umich.edu/dlps/mbonn-saunders.html.

Durfee, Edward, Daniel Kiskis, and William Birmingham. "The Agent Architecture of the University of Michigan Digital Library." *IEE-Proceedings-Software-Engineering.* vol.144, no.1; Feb. 1997; p.61-71.

Finholt, Thomas. "Evaluation of Electronic Work: Research on Collaboratories at the University of Michigan." *SIGOIS-Bulletin.* vol.16, no.2; Dec. 1995; p.49-51.

Guthrie, Kevin. "JSTOR: From Project to Independent Organization," *D-Lib*, July-August 1997. http://www.dlib.org/dlib/july97/07guthrie.html.

Information and People: A Campus Dialogue on the Challenges of Electronic Information. Final Report of the Information Symposium. Ann Arbor: School of Information and Library Studies, March, 1991.

Lougee, Wendy. School for Scanning Presentation. need citation.

Mackie-Mason, Jeffrey and Alexandra Jankovich. "PEAK: Pricing Electronic Access to Knowledge at the University of Michigan; presented at the Elsevier Electronic Subscriptions conference, October 1996." *Library Acquisitions*, 21:281-95 Fall '97 See also http://www-personal.umich.edu/~jmm/papers/PEAK/.

Powell, Christina Kelleher and Nigel Kerr. "SGML Creation and Delivery: The Humanities Text Initiative." *D-Lib Magazine*, July/August 1997. See http://www.dlib.org/dlib/july97/humanities/07powell.html.

Price-Wilkin, John. "Just-in-time Conversion, Just-in-case Collections: Effectively Leveraging Rich Document Formats for the WWW." *D-Lib Magazine*, May 1997. See http://www.dlib.org/dlib/may97/michigan/05pricewilkin.html.

Price-Wilkin, John. "Text Files in Libraries: Present Foundations and Future Directions," *Library Hi Tech*, Consecutive Issue 35, (1991)7-44.

Shaw, Elizabeth and Sarr Blumson. "Making of America; Online Searching and Page Presentation at the University of Michigan." *D-Lib Magazine*, July/August 1997. See http://www.dlib.org/dlib/july97/america/07shaw.html.

Stephenson, Christie, and Patricia McClung. *MESL: Delivering Digital Images. Cultural Heritage Resources for Education.* Los Angeles: The Getty Information Institute, 1998.

Waters, Don. "A Working Definition of Digital Library," on the Digital Library Federation Web site. http://www.clir.org/diglib/dldefinition.htm.

York, Grace. "A Facelift for Tradition: Mainstreaming Government Information on the Internet," in *Proceedings of the 3rd Annual Federal Depository Library Conference*, April 20-22, 1994, pp. 133-139.

York, Grace. "Out of the Basement: The Internet and Document Public Services," in *Proceedings of the 7th Annual Federal Depository Library Conference*, April 20-23, 1998, pp. 170-176.

FIGURE 1

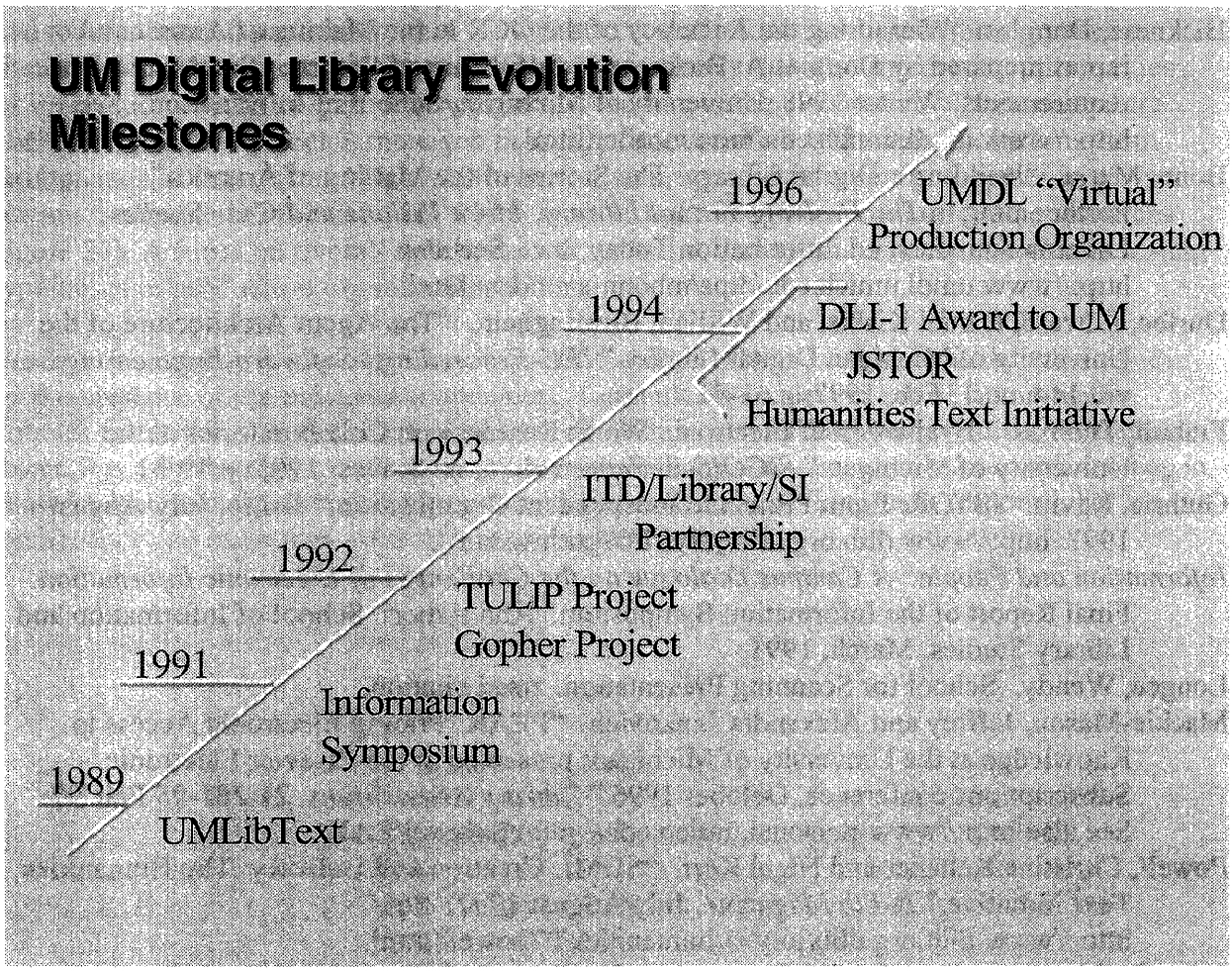

UM Digital Library Evolution
Milestones
1996
UMDL "Virtual"
Production Organization
1994
DLI-1 Award to UM
JSTOR
Humanities Text Initiative
1993
ITD/Library/SI
Partnership
1992
TULIP Project
Gopher Project
1991
Information
Symposium
1989
UMLibText

FIGURE 2: DLPS ORGANIZATION CHART

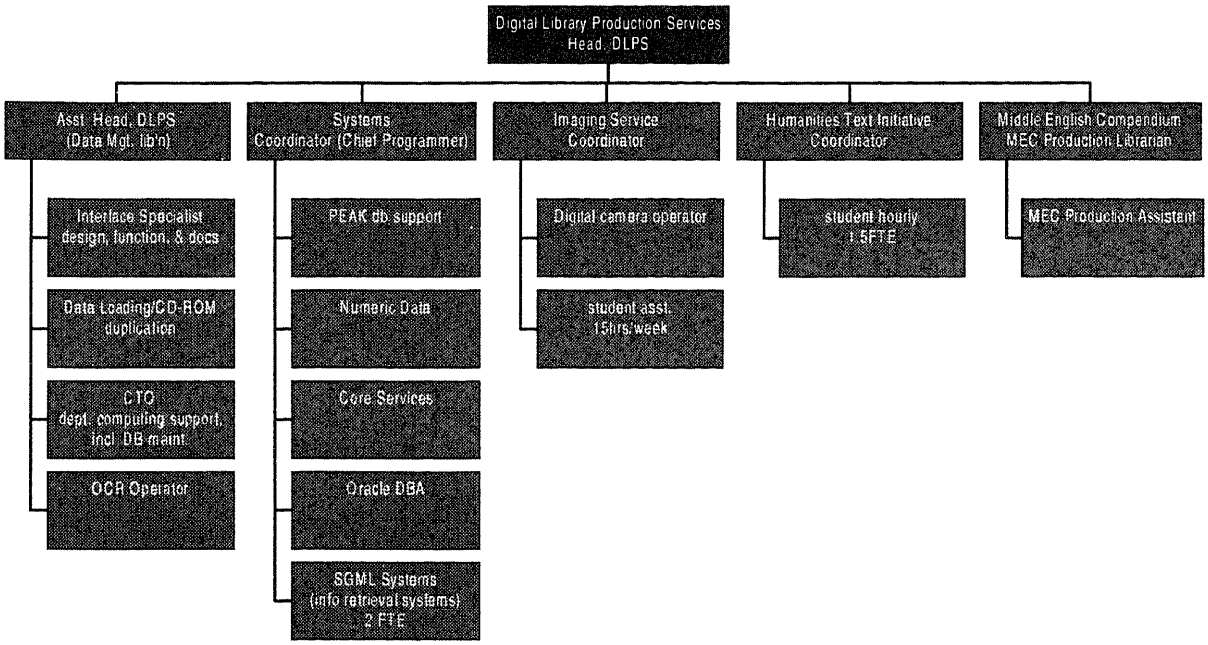
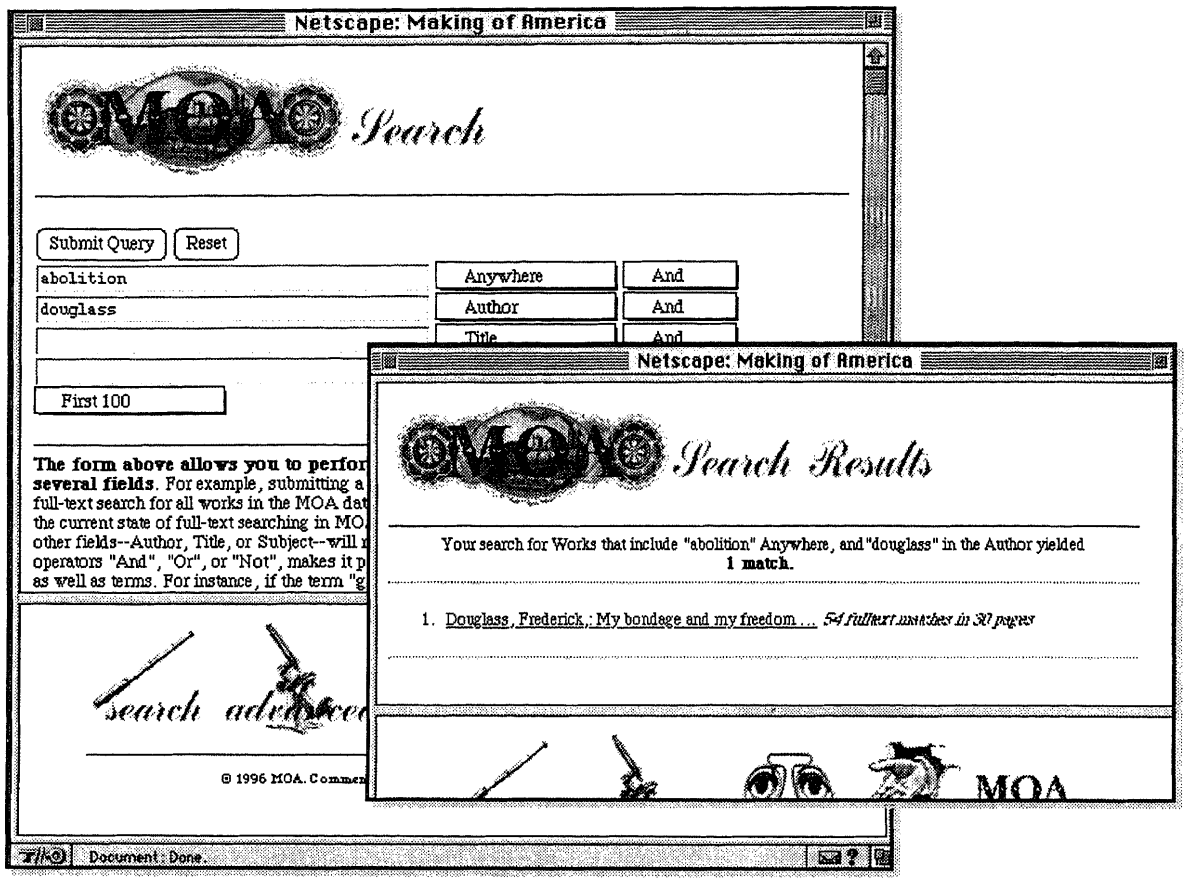
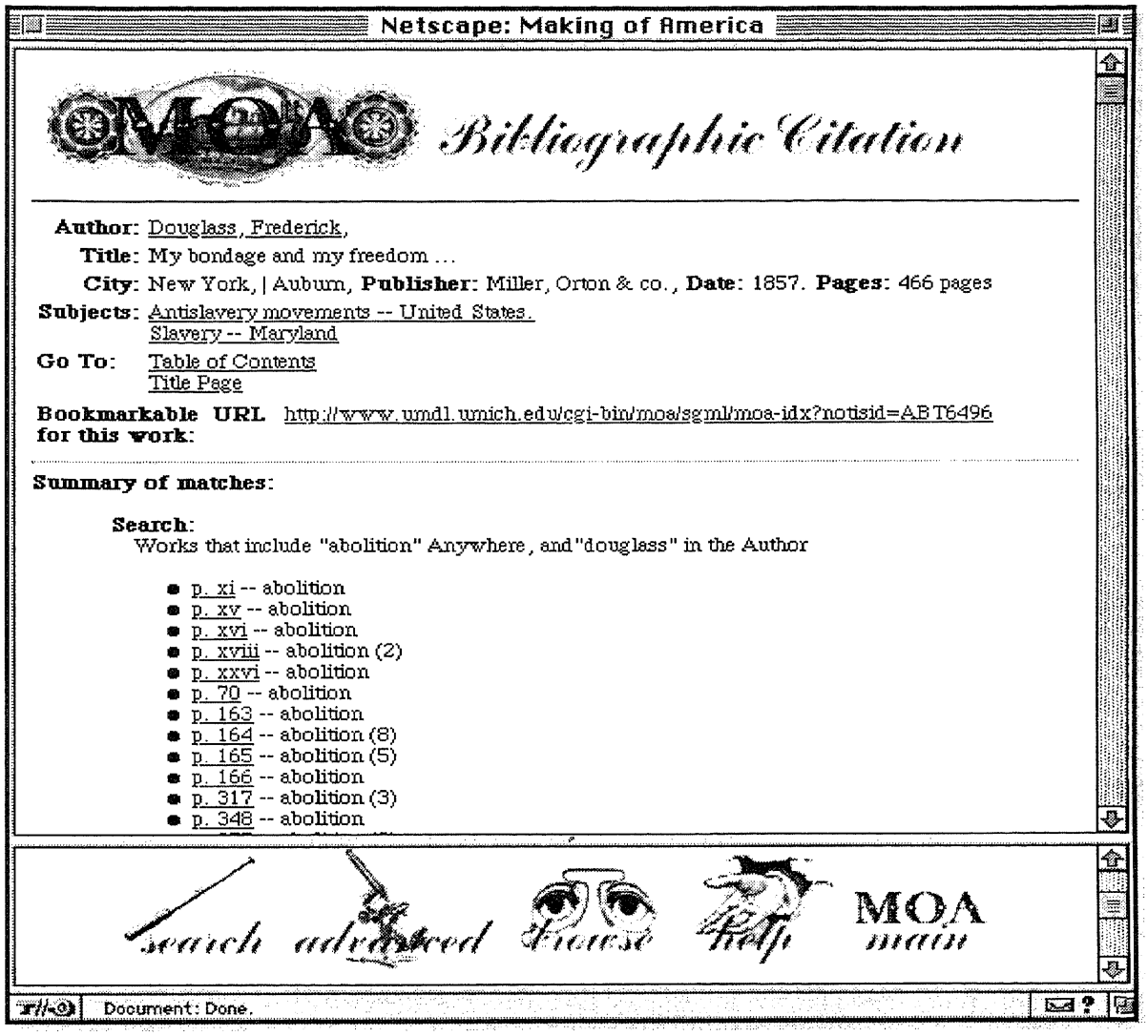


FIGURE 3: MAKING OF AMERICA—SEARCH INTERFACE AND SAMPLE RESULT SCREEN



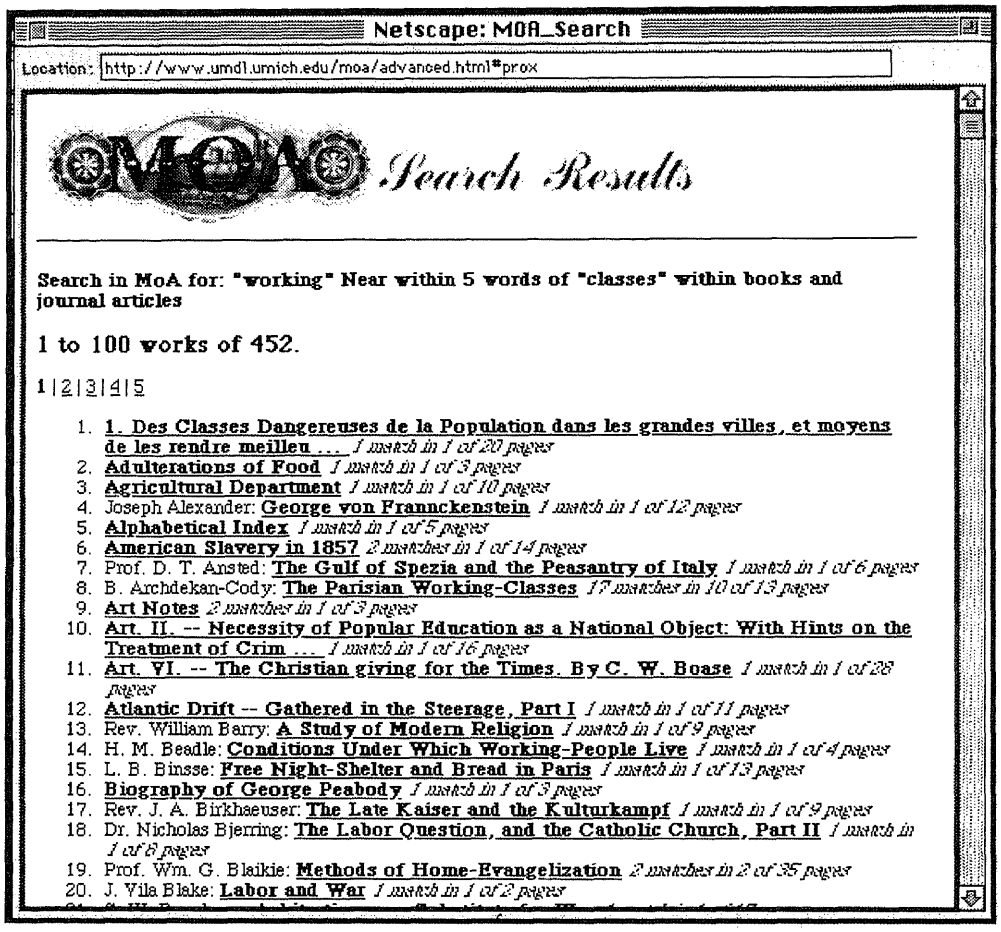
The Making of America system offers a variety of types of searches, including the advanced Boolean search seen above.

FIGURE 4: MAKING OF AMERICA—LOCATION OF RESULTS WITHIN SINGLE TEXT



MOA results (within a selected text) bring the user to the page or pages containing results, and give the user information on the distribution of results within those pages.

FIGURE 5: MAKING OF AMERICA—SAMPLE RESULTS SCREEN (RELEVANCE FEEDBACK)



Results from MOA provide important relevance feedback. Weighting in volumes as large and various as the MOA volumes is particularly problematic. Instead, results are presented to the user with important information on the number of occurrences, the total number of pages, and the number of pages containing matches.

FIGURE 6: MAKING OF AMERICA—AUTOMATICALLY GENERATED GIF PAGE DISPLAY AND OTHER DISPLAY RESOLUTIONS

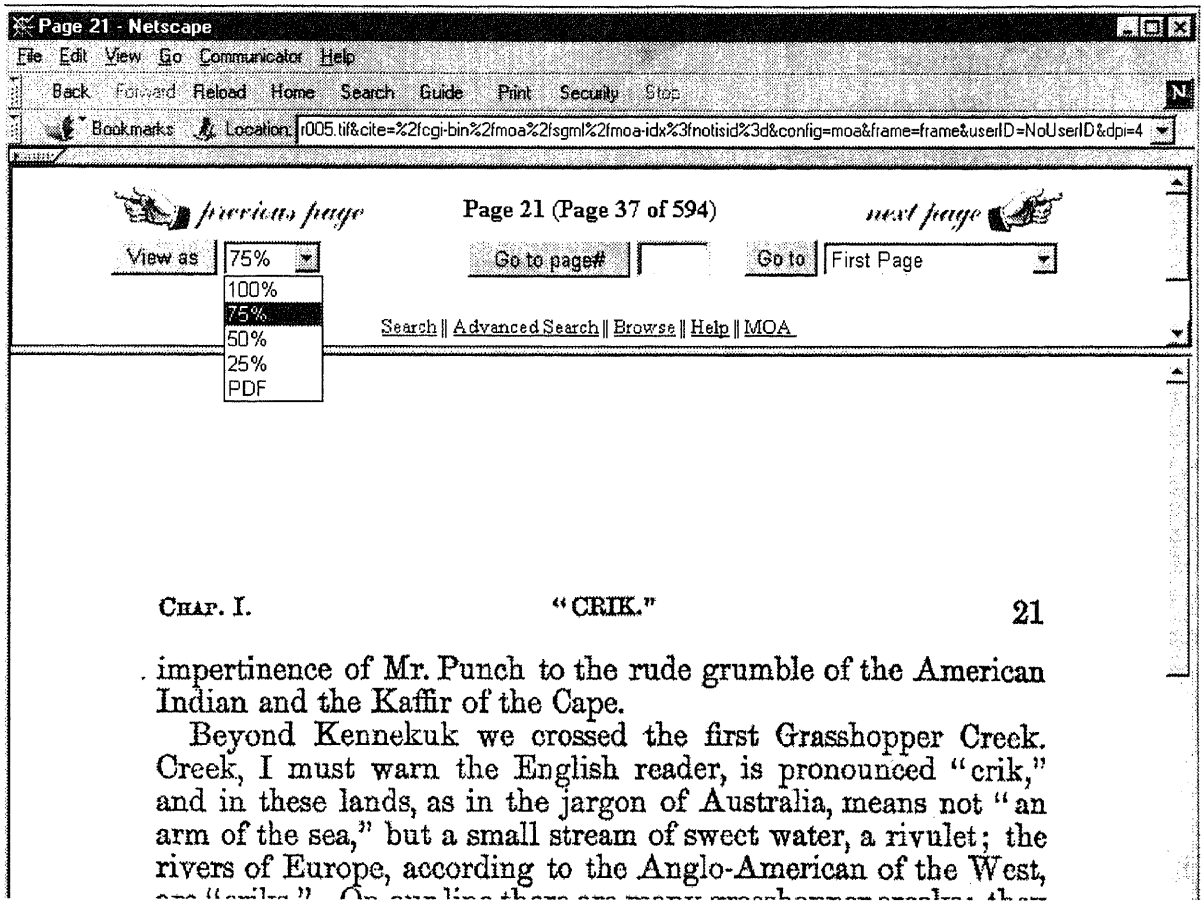
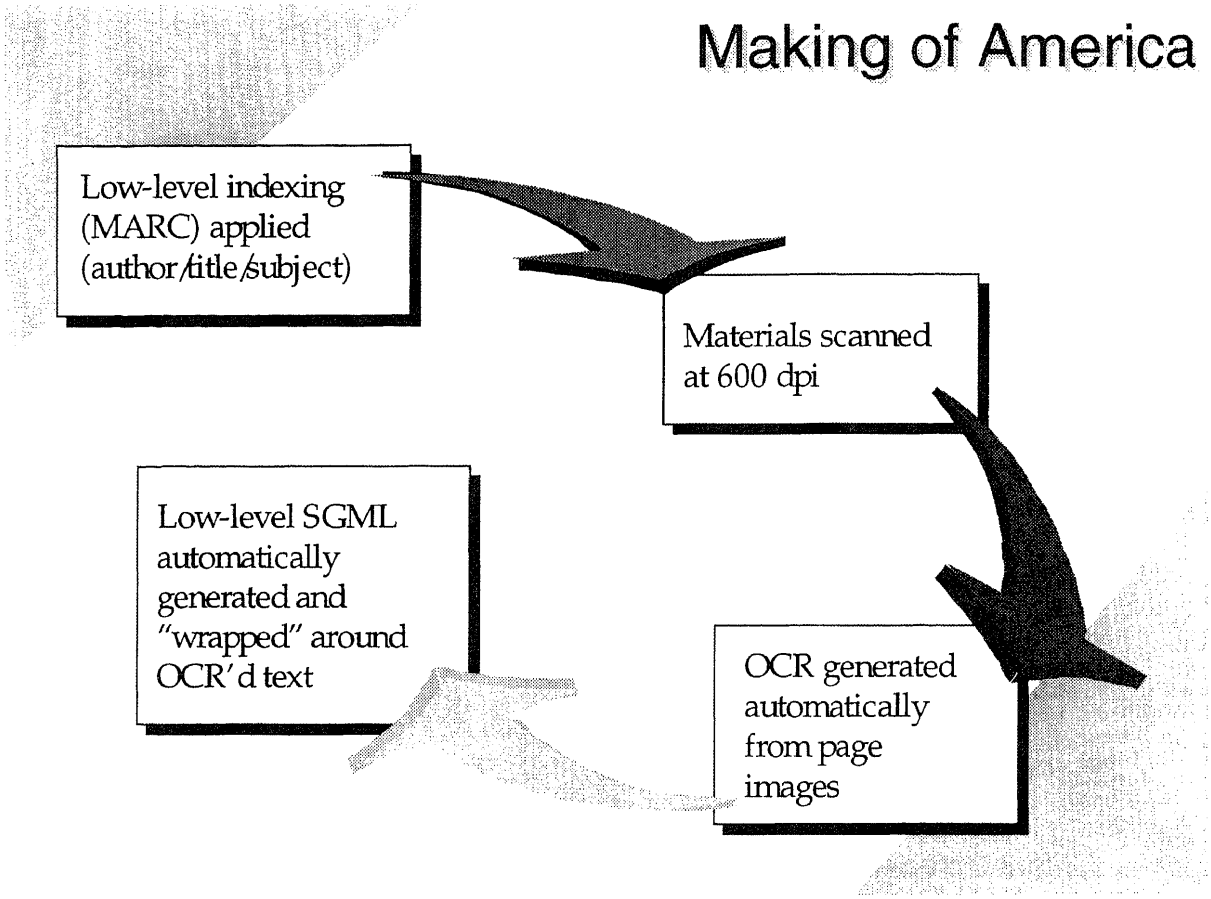
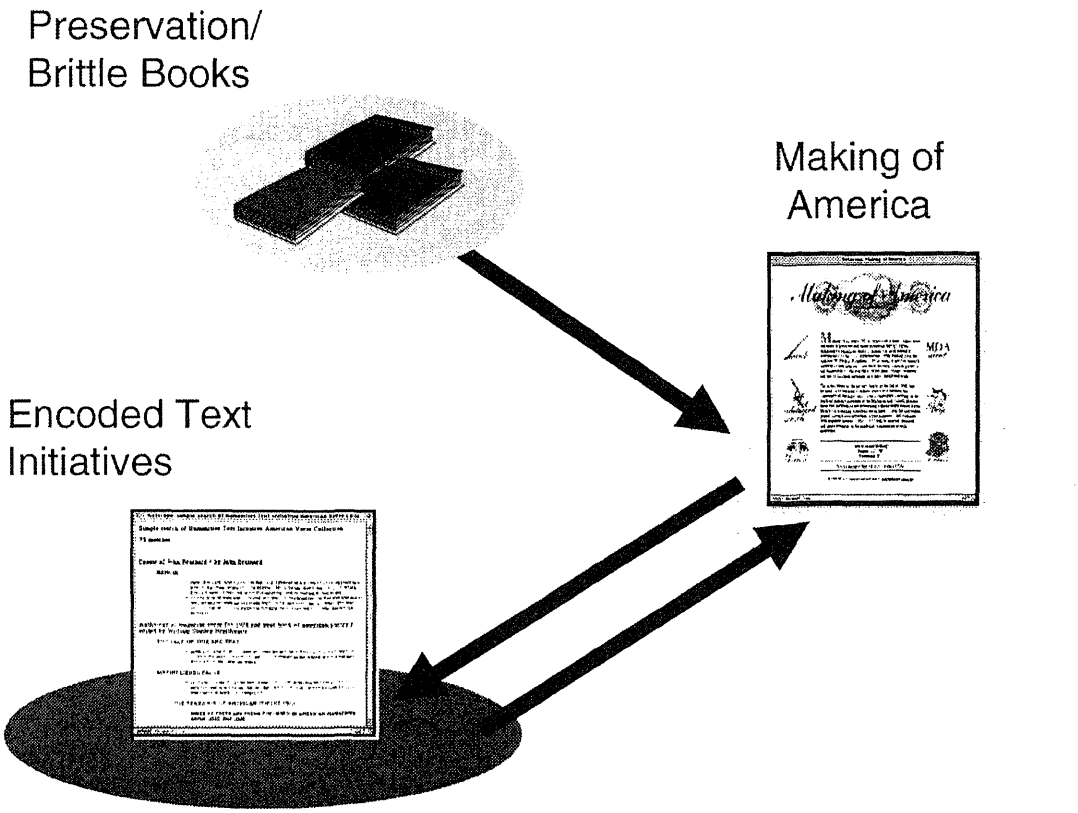


FIGURE 7: PRODUCTION MODEL FOR MAKING OF AMERICA



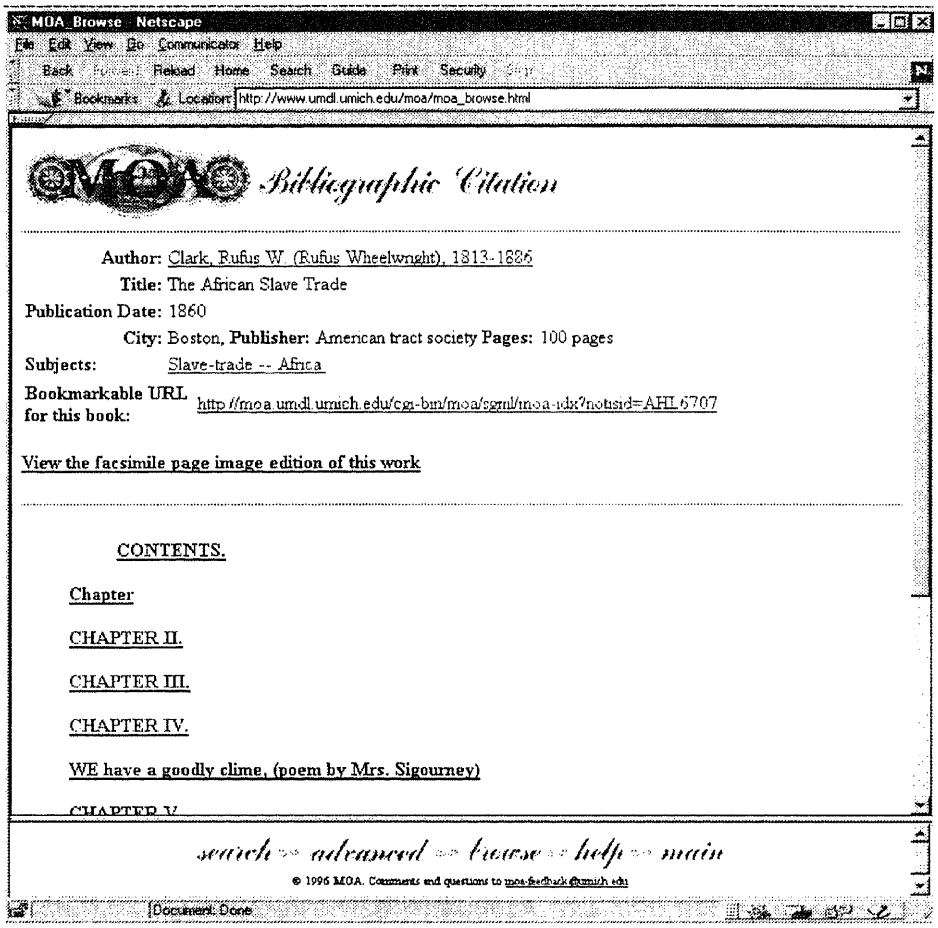
All MOA processing begins with MARC records and basic collation information derived in the process of preparing the volumes (e.g., number of pages, page numbering, etc.). After routine scanning, typically by vendors, OCR is generated at Michigan. This OCR, along with the MARC record and collation information, is bound together by a rudimentary SGML consistent with the TEI Guidelines.

FIGURE 8: MAKING OF AMERICA—INTEGRATION OF ENCODED TEXTS IN PAGE IMAGES



All MOA materials begin with evaluation and treatment by the Preservation Department. The typical MOA object, page-image based resource, is created inexpensively and quickly. This growing body of materials then serves as a collection source for DLPS's encoded text initiatives (in the Humanities Text Initiative), where individual volumes may be selected for refinement of encoding, as well as correction of OCR. The resulting volume replaces the automatically generated material in MOA.

FIGURE 9: SAMPLE "TABLE OF CONTENTS" DISPLAY FROM ENCODED TEXT IN MOA



A MOA volume that has been fully encoded and whose OCR has been corrected may subsequently be viewed as an HTML document (with the SGML text being used to generate the HTML at the point of use). This more functional version is particularly useful for vision-impaired users of the Web, as well as for textual analysis, but importantly speeds transfer times for ordinary users and allows more effective searching and reading of the text.

FIGURE 10: METADATA MAPPING IN DLPS IMAGE SERVICES

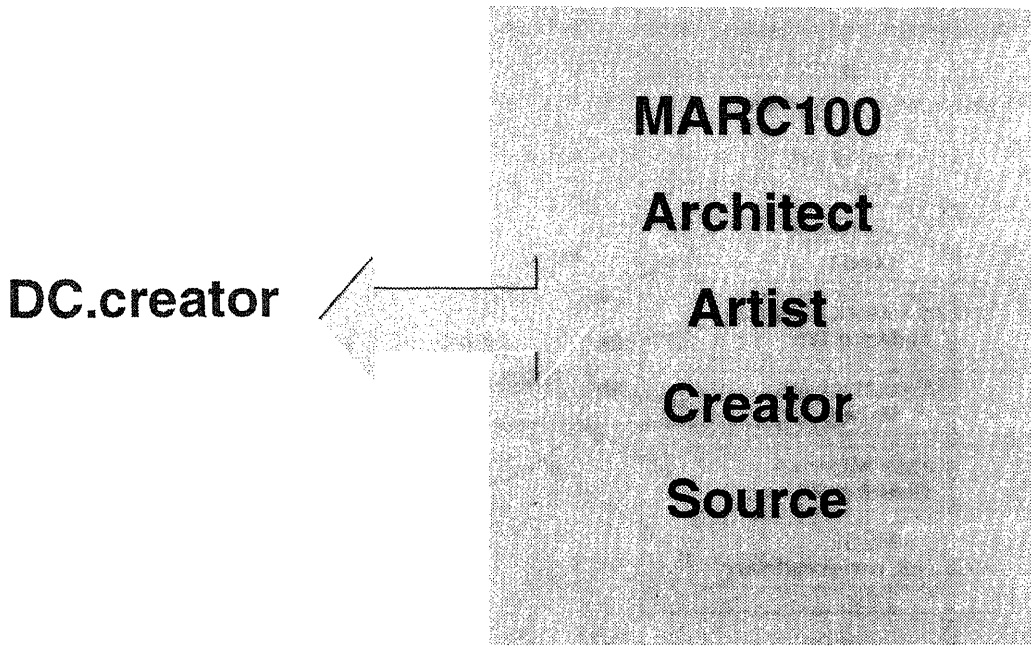
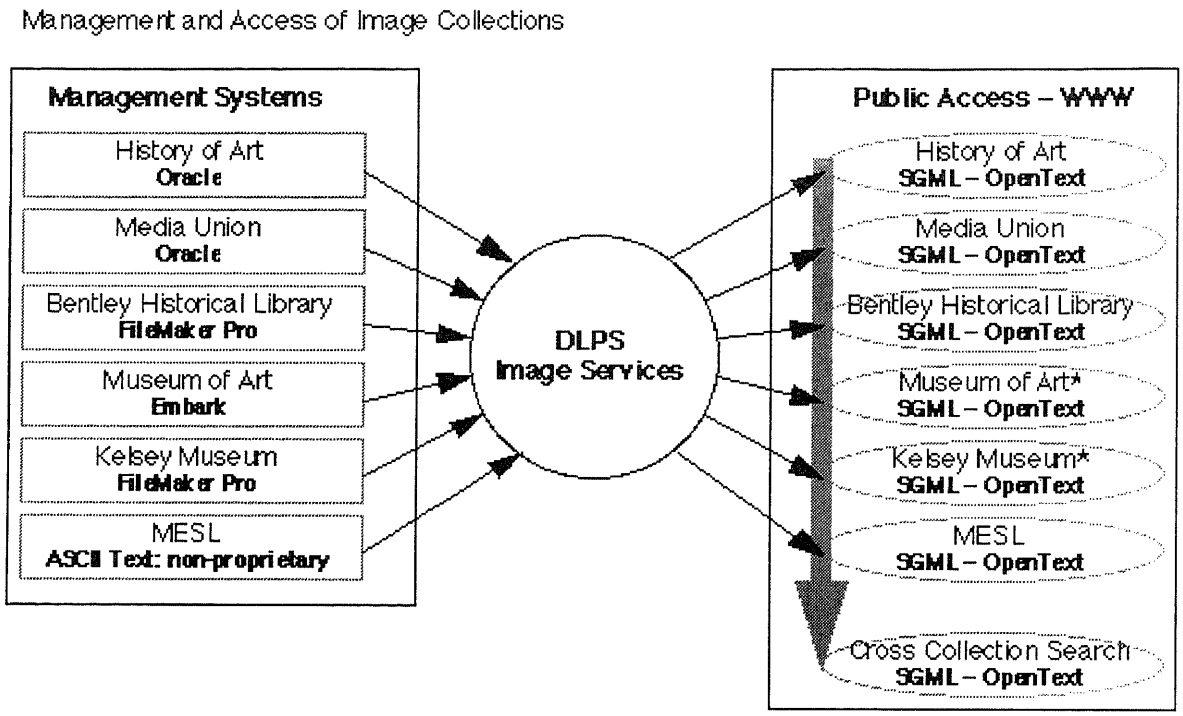


Image Services uses a metadata mapping strategy that links the field labels of the contributing institution to Dublin Core fields. Both the original fields and the Dublin Core fields are retained in the system, enabling users to approach the system as if it were a unified database, or the unmodified database of the contributing organization.

FIGURE 11: DATABASE MERGING IN DLPS IMAGE SERVICES



The approach taken by DLPS Image Services brings together databases from a diverse group of systems intended primarily for management of data. By separating "management" and "access" strategies in this way, we are able to create more flexible and more responsive access systems that are unburdened by the needs associated with data management.

FIGURE 12: SAMPLE IMAGE SERVICES SEARCH INTERFACE

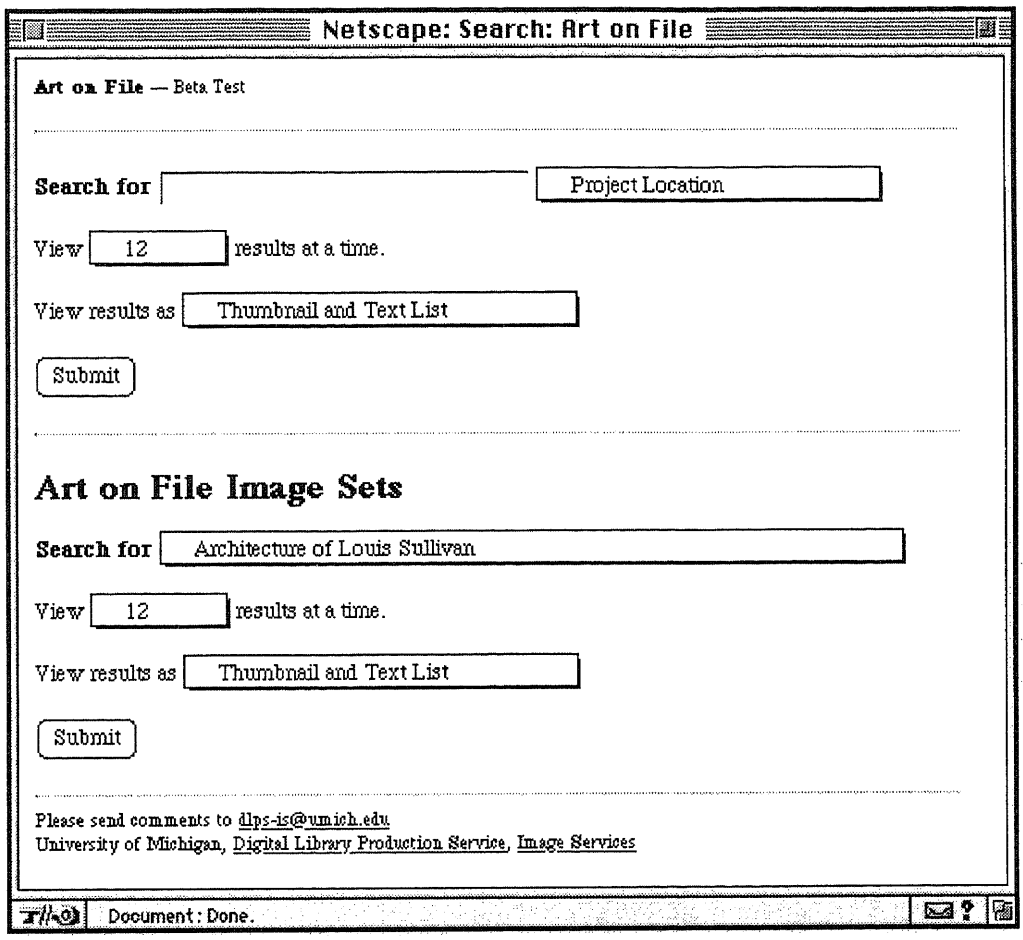
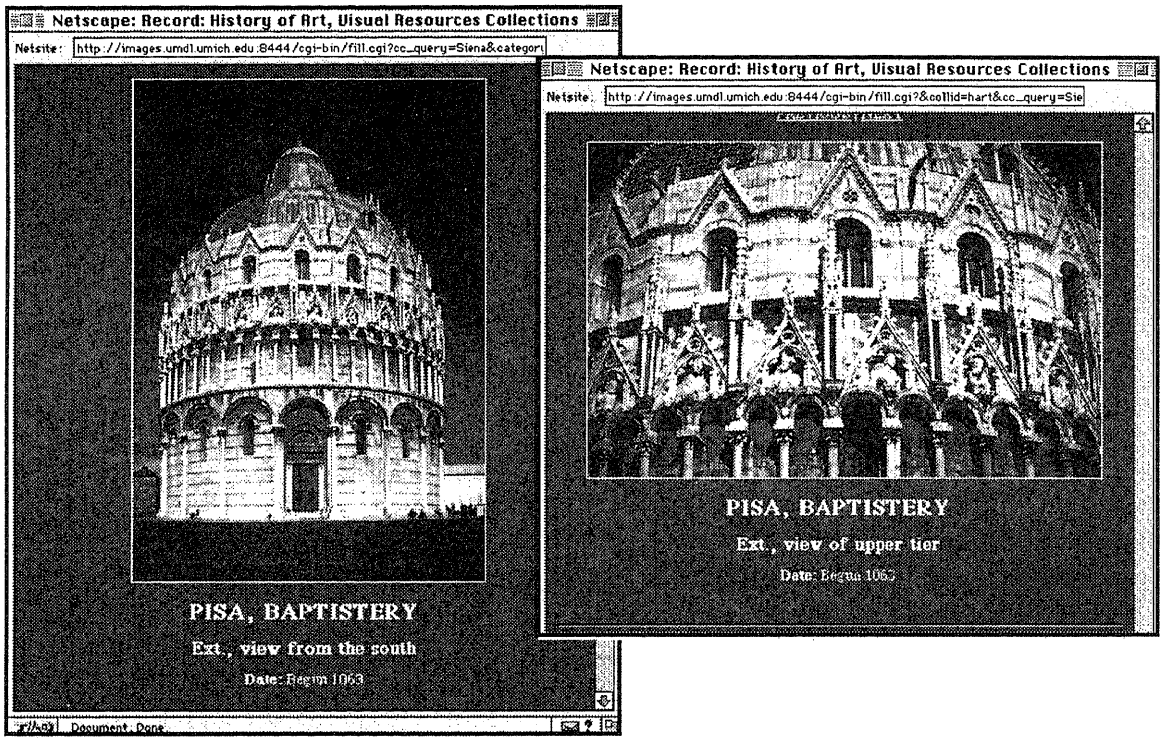


FIGURE 13: SAMPLE IMAGE SERVICES "SLIDE SHOW" VIEW OF A COLLECTION



This view (above) of the Image Services database is from the same database as that which generates typical search results. The difference in presentation—in this case a slide show, with pre-selected views and order—is determined by a "template" whose layout communicates with the image database.

FIGURE 14: COMPARISON FUNCTIONALITY IN IMAGE SERVICES

Another template (cf. the "slide show" template above) available in the Image Services system is the "comparison" template. It allows individually selected search results to be presented in panels, side by side.

FIGURE 15: "PAN AND ZOOM" FEATURE IN IMAGE SERVICES

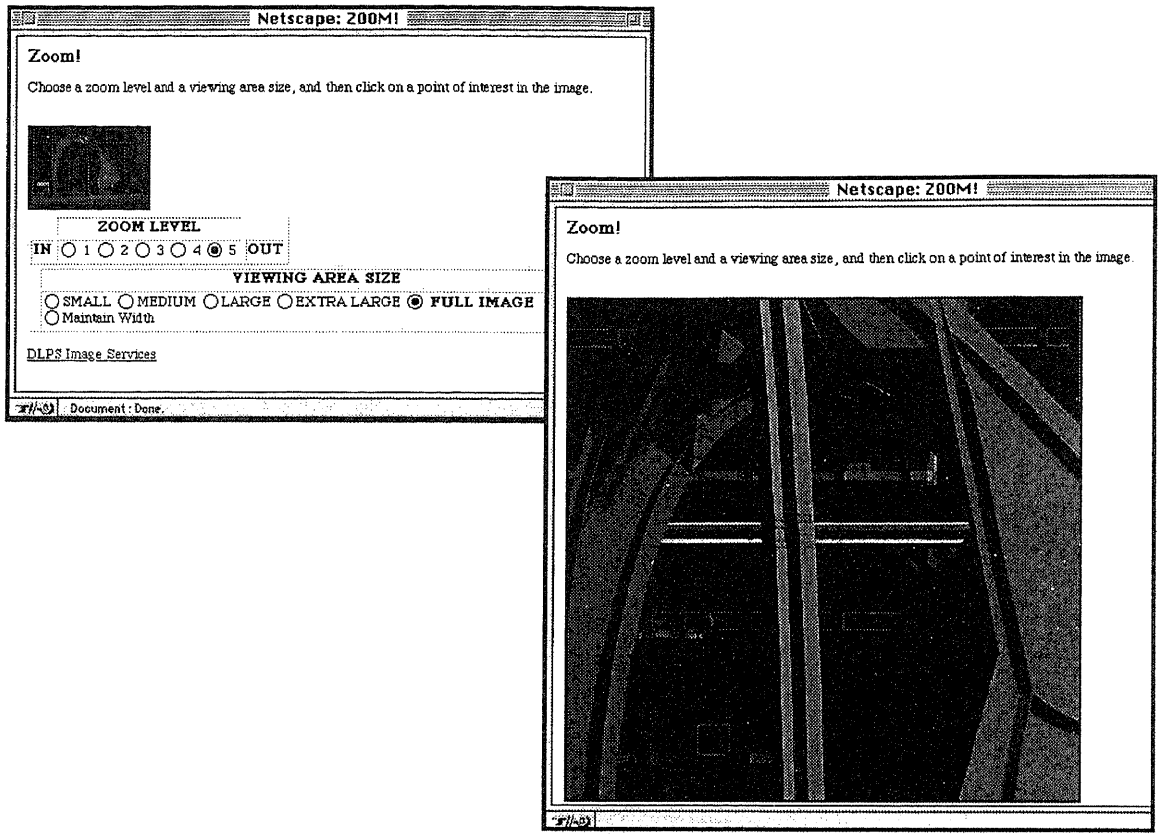


FIGURE 16: AUTHENTICATION SCREEN FROM PEAK SYSTEM

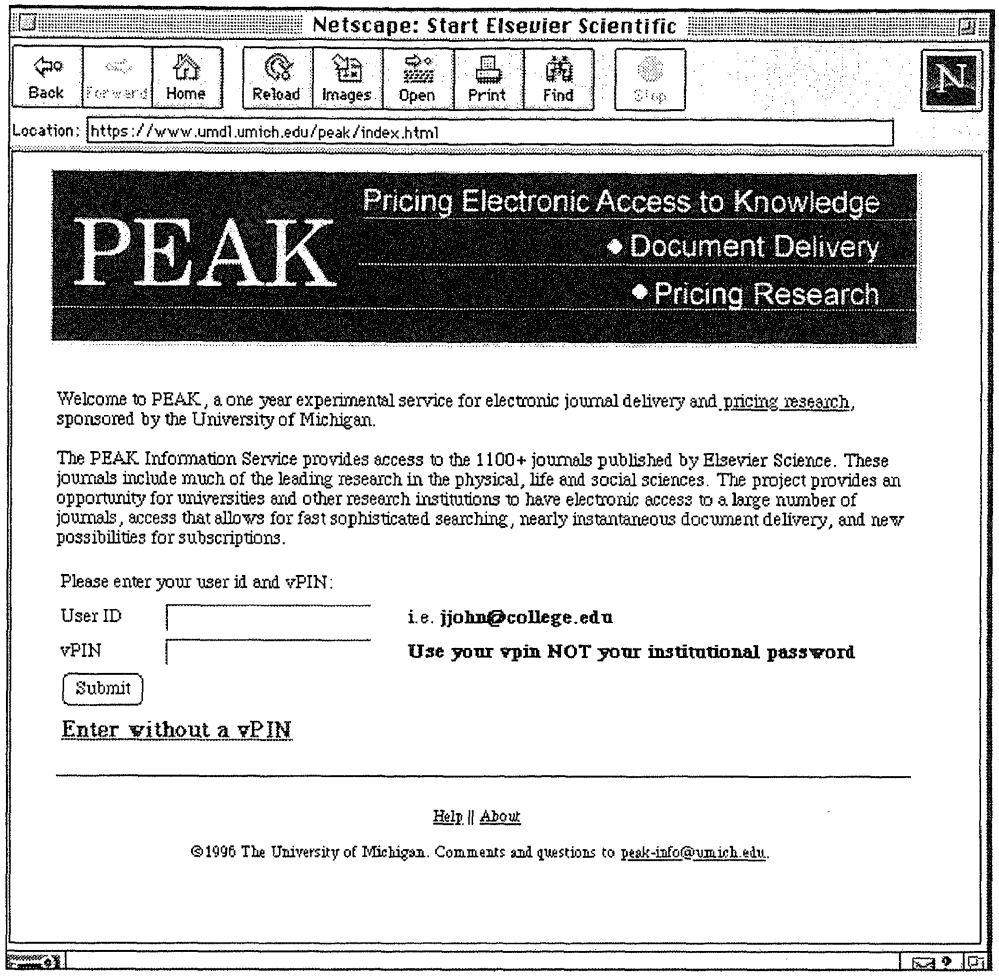


FIGURE 17: PEAK—SIMPLE SEARCH INTERFACE AND SAMPLE RESULTS SCREEN

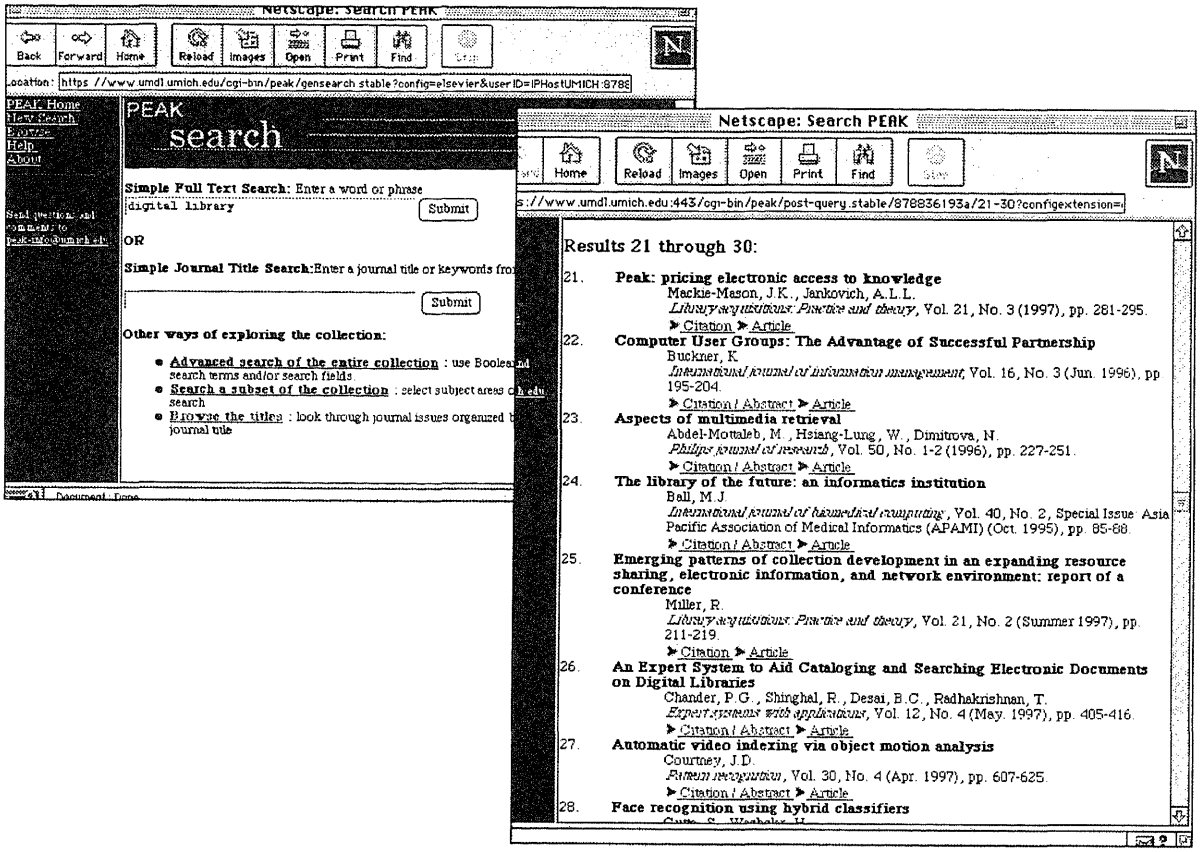


FIGURE 18: PEAK—ADVANCED SEARCH INTERFACE AND SAMPLE BROWSE INTERFACE

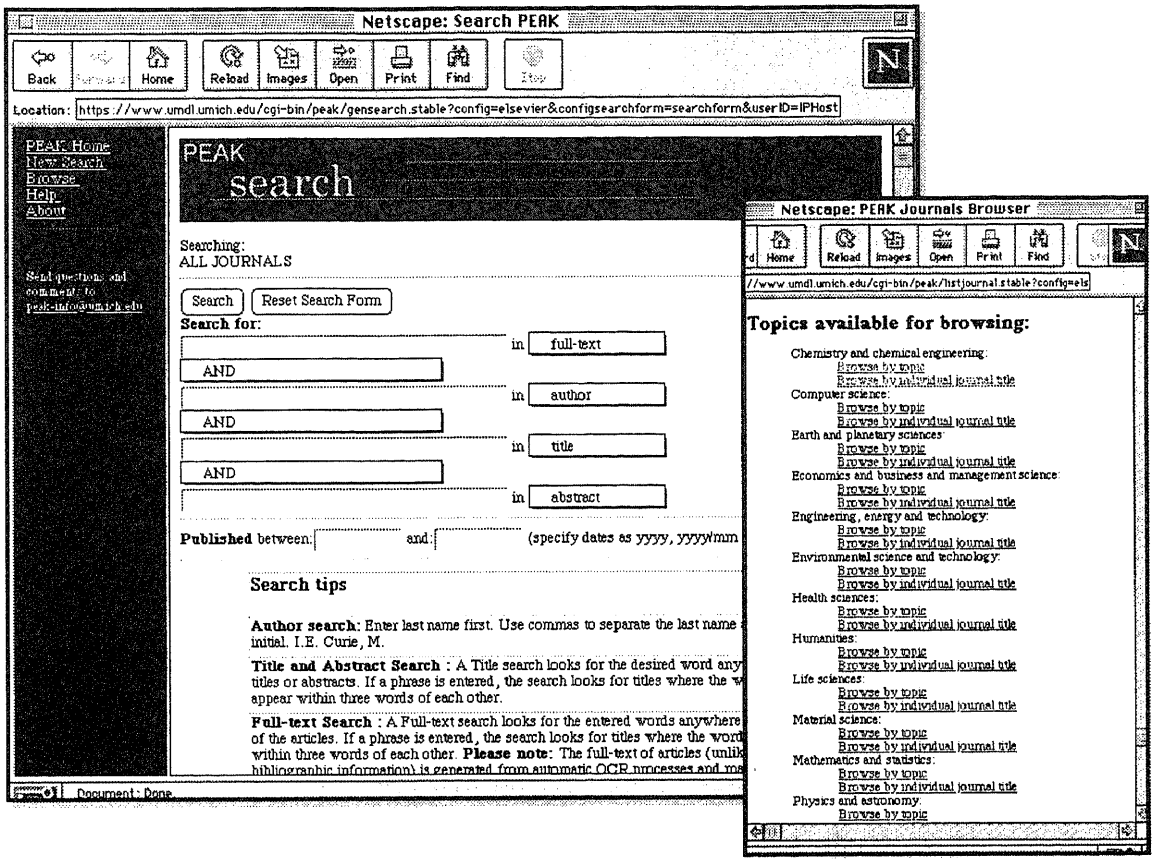


FIGURE 19: PEAK—AUTOMATICALLY GENERATED PAGE DISPLAY & PRINT OPTIONS

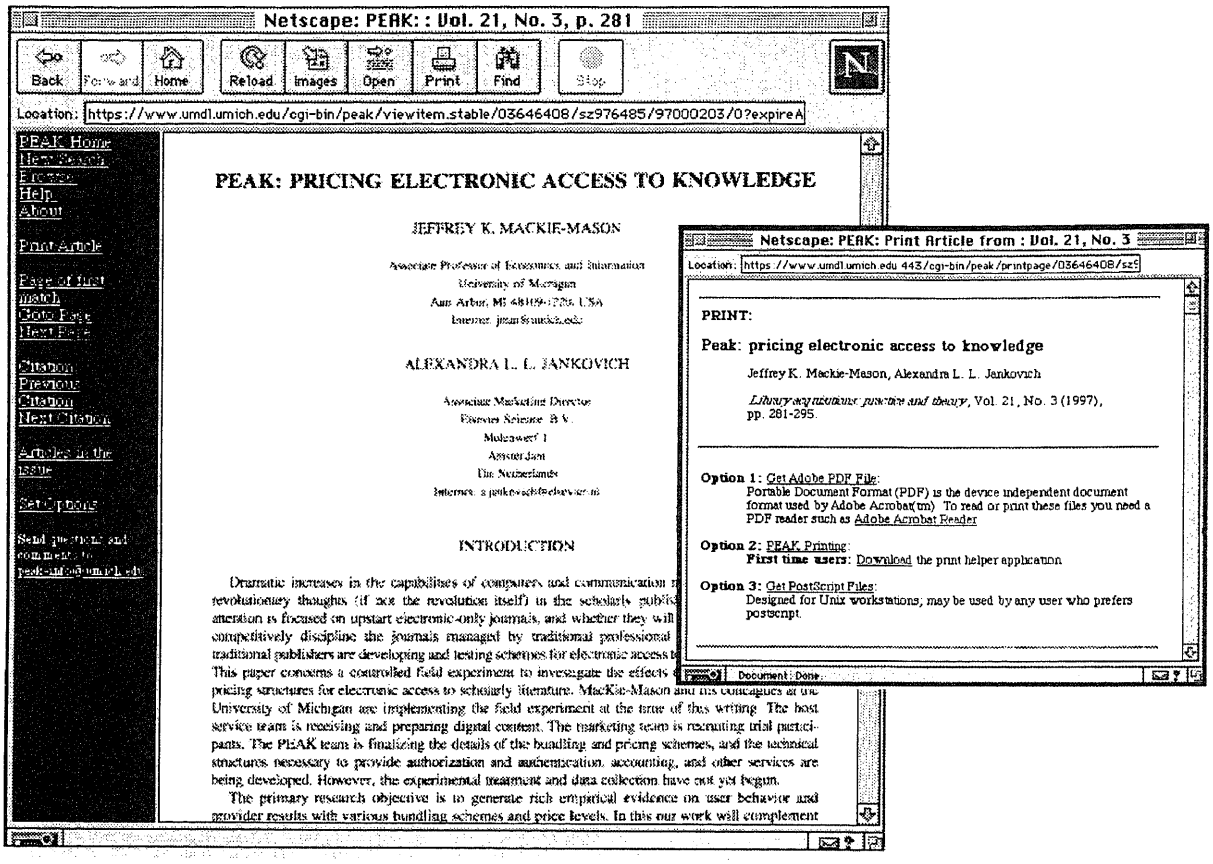


FIGURE 20

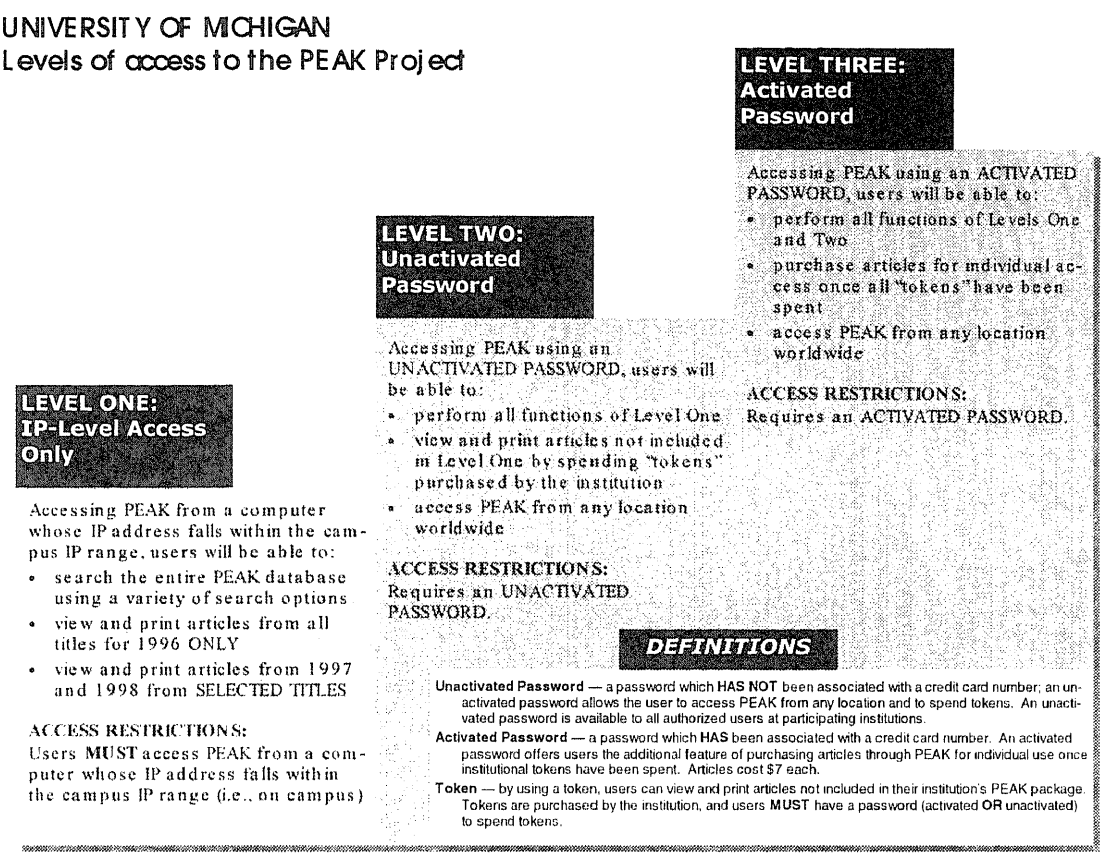
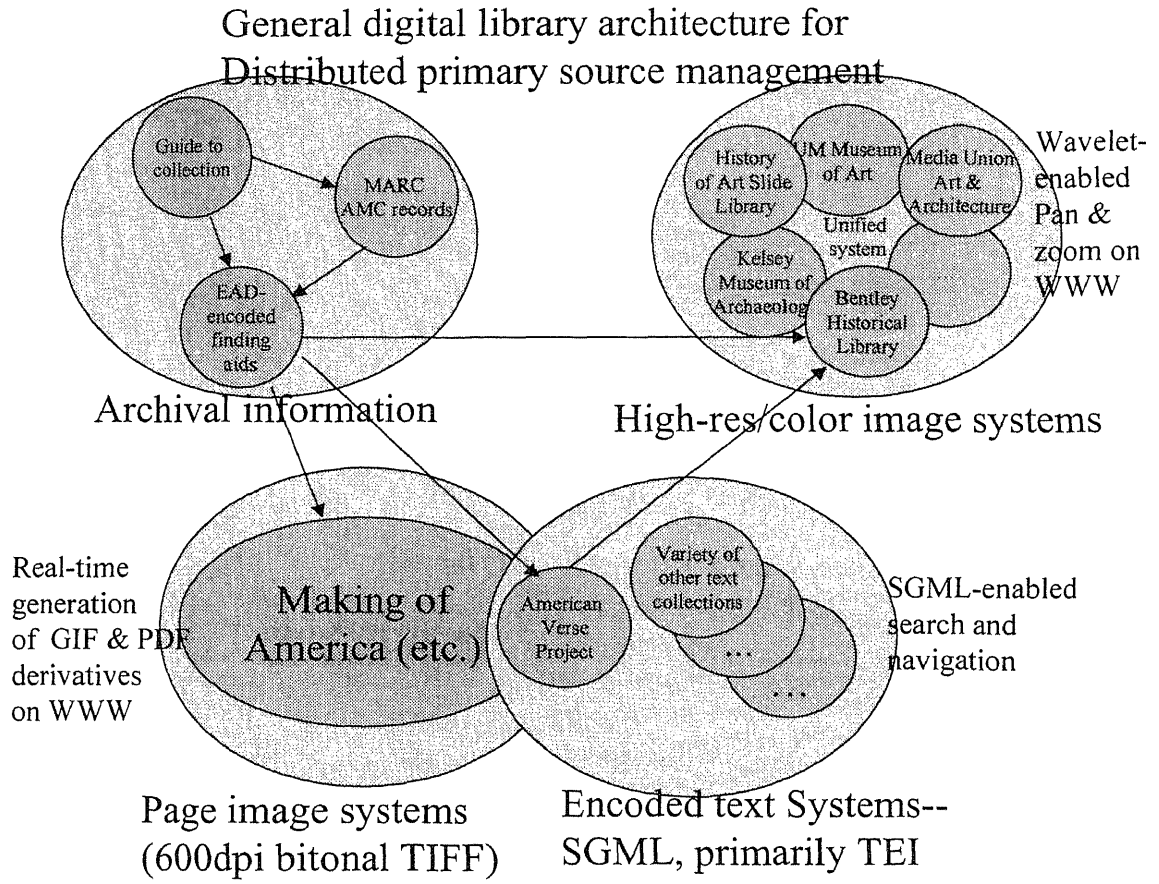

UNIVERSITY OF MICHIGAN
Levels of access to the PEAK Project
LEVEL ONE: IP-Level Access Only
Accessing PEAK from a computer whose IP address falls within the campus IP range, users will be able to:
• search the entire PEAK database using a variety of search options
• view and print articles from all titles for 1996 ONLY
• view and print articles from 1997 and 1998 from SELECTED TITLES
ACCESS RESTRICTIONS:
Users MUST access PEAK from a computer whose IP address falls within the campus IP range (i.e., on campus)
LEVEL TWO: Unactivated Password
Accessing PEAK using an UNACTIVATED PASSWORD, users will be able to:
• perform all functions of Level One
• view and print articles not included in Level One by spending "tokens" purchased by the institution
• access PEAK from any location worldwide
ACCESS RESTRICTIONS:
Requires an UNACTIVATED PASSWORD.
LEVEL THREE: Activated Password
Accessing PEAK using an ACTIVATED PASSWORD, users will be able to:
• perform all functions of Levels One and Two
• purchase articles for individual access once all "tokens" have been spent
• access PEAK from any location worldwide
ACCESS RESTRICTIONS:
Requires an ACTIVATED PASSWORD.
DEFINITIONS
Unactivated Password — a password which HAS NOT been associated with a credit card number; an unactivated password allows the user to access PEAK from any location and to spend tokens. An unactivated password is available to all authorized users at participating institutions.
Activated Password — a password which HAS been associated with a credit card number. An activated password offers users the additional feature of purchasing articles through PEAK for individual use once institutional tokens have been spent. Articles cost $7 each.
Token — by using a token, users can view and print articles not included in their institution's PEAK package. Tokens are purchased by the institution, and users MUST have a password (activated OR unactivated) to spend tokens.

FIGURE 21: INTEGRATION OF DISTRIBUTED AND MULTI-TYPE OBJECTS IN DLPS



The different systems for retrieval in DLPS are joined through commonly understood (and, we hope soon, published) methods for inter-system communication. Each system is designed to support a *type* of material, and consequently may will entirely unique retrieval, display, and management tools. Still, the resources are integrated in important ways. In this view illustrating the organization of primary source materials from the Bentley Historical Library (above), descriptive information from the Bentley is linked to continuous tone images (e.g., images of photographs), bitonal page images, and encoded texts, with each type of information supporting links to the others.