WORKSHOP ON ELECTRONIC TEXTS

PROCEEDINGS

Edited by James Daly

9-10 June 1992

Library of Congress

Washington, D.C.

Supported by a Grant from the David and Lucile Packard Foundation

*** *** *** ****** *** *** ***

TABLE OF CONTENTS

Discussion

Lynne K. Personius, Cornell mathematics books

Discussion

Session III. Distribution, Networks, and Networking:

Options for Dissemination

Robert G. Zich (Moderator)

Clifford A. Lynch

Discussion

Howard Besser

Discussion

Ronald L. Larsen

Edwin B. Brownrigg

Discussion

Session IV. Image Capture, Text Capture, Overview of Text and

Image Storage Formats

William L. Hooton (Moderator)

A) Principal Methods for Image Capture of Text:

direct scanning, use of microform

Anne R. Kenney

Pamela Q.J. Andre

Judith A. Zidar

Donald J. Waters

Discussion

B) Special Problems: bound volumes, conservation,

reproducing printed halftones

*** *** *** ****** *** *** ***

Acknowledgements

I would like to thank Carl Fleischhauer and Prosser Gifford for the opportunity to learn about areas of human activity unknown to me a scant ten months ago, and the David and Lucile Packard Foundation for supporting that opportunity. The help given by others is acknowledged on a separate page.

19 October 1992

*** *** *** ****** *** *** ***

INTRODUCTION

The Workshop on Electronic Texts (1) drew together representatives of various projects and interest groups to compare ideas, beliefs, experiences, and, in particular, methods of placing and presenting historical textual materials in computerized form. Most attendees gained much in insight and outlook from the event. But the assembly did not form a new nation, or, to put it another way, the diversity of projects and interests was too great to draw the representatives into a cohesive, action-oriented body.(2)

Everyone attending the Workshop shared an interest in preserving and providing access to historical texts. But within this broad field the attendees represented a variety of formal, informal, figurative, and literal groups, with many individuals belonging to more than one. These groups may be defined roughly according to the following topics or activities:

* Imaging
* Searchable coded texts
* National and international computer networks
* CD-ROM production and dissemination
* Methods and technology for converting older paper materials into electronic form
* Study of the use of digital materials by scholars and others

This summary is arranged thematically and does not follow the actual sequence of presentations.

NOTES:

(1) In this document, the phrase electronic text is used to mean any computerized reproduction or version of a document, book, article, or manuscript (including images), and not merely a machine-readable or machine-searchable text.

(2) The Workshop was held at the Library of Congress on 9-10 June 1992, with funding from the David and Lucile Packard Foundation. The document that follows represents a summary of the presentations made at the Workshop and was compiled by James DALY. This introduction was written by DALY and Carl FLEISCHHAUER.

PRESERVATION AND IMAGING

Preservation, as that term is used by archivists,(3) was most explicitly discussed in the context of imaging. Anne KENNEY and Lynne PERSONIUS explained how the concept of a faithful copy and the user-friendliness of the traditional book have guided their project at Cornell University.(4) Although interested in computerized dissemination, participants in the Cornell project are creating digital image sets of older books in the public domain as a source for a fresh paper facsimile or, in a future

phase, microfilm. The books returned to the library shelves are high-quality and useful replacements on acid-free paper that should last a long time. To date, the Cornell project has placed little or no emphasis on creating searchable texts; one would not be surprised to find that the project participants view such texts as new editions, and thus not as faithful reproductions.

In her talk on preservation, Patricia BATTIN struck an ecumenical and flexible note as she endorsed the creation and dissemination of a variety of types of digital copies. Do not be too narrow in defining what counts as a preservation element, BATTIN counseled; for the present, at least, digital copies made with preservation in mind cannot be as narrowly standardized as, say, microfilm copies with the same objective. Setting standards precipitously can inhibit creativity, but delay can result in chaos, she advised.

In part, BATTIN's position reflected the unsettled nature of image-format standards, and attendees could hear echoes of this unsettledness in the comments of various speakers. For example, Jean BARONAS reviewed the status of several formal standards moving through committees of experts; and Clifford LYNCH encouraged the use of a new guideline for transmitting document images on Internet. Testimony from participants in the National Agricultural Library's (NAL) Text Digitization Program and LC's American Memory project highlighted some of the challenges to the actual creation or interchange of images, including difficulties in converting preservation microfilm to digital form. Donald WATERS reported on the progress of a master plan for a project at Yale University to convert

books on microfilm to digital image sets, Project Open Book (POB).

The Workshop offered rather less of an imaging practicum than planned, but "how-to" hints emerge at various points, for example, throughout KENNEY's presentation and in the discussion of arcana such as thresholding and dithering offered by George THOMA and FLEISCHHAUER.

NOTES:

(3) Although there is a sense in which any reproductions of historical materials preserve the human record, specialists in the field have developed particular guidelines for the creation of acceptable preservation copies.

(4) Titles and affiliations of presenters are given at the beginning of their respective talks and in the Directory of Participants (Appendix III).

THE MACHINE-READABLE TEXT: MARKUP AND USE

The sections of the Workshop that dealt with machine-readable text tended to be more concerned with access and use than with preservation, at least in the narrow technical sense. Michael SPERBERG-McQUEEN made a forceful presentation on the Text Encoding Initiative's (TEI) implementation of the Standard Generalized Markup Language (SGML). His ideas were echoed by Susan HOCKEY, Elli MYLONAS, and Stuart WEIBEL. While the

presentations made by the TEI advocates contained no practicum, their discussion focused on the value of the finished product, what the European Community calls reusability, but what may also be termed durability. They argued that marking up--that is, coding--a text in a well-conceived way will permit it to be moved from one computer environment to another, as well as to be used by various users. Two kinds of markup were distinguished: 1) procedural markup, which describes the features of a text (e.g., dots on a page), and 2) descriptive markup, which describes the structure or elements of a document (e.g., chapters, paragraphs, and front matter).

The TEI proponents emphasized the importance of texts to scholarship. They explained how heavily coded (and thus analyzed and annotated) texts can underlie research, play a role in scholarly communication, and facilitate classroom teaching. SPERBERG-McQUEEN reminded listeners that a written or printed item (e.g., a particular edition of a book) is merely a representation of the abstraction we call a text. To concern ourselves with faithfully reproducing a printed instance of the text, SPERBERG-McQUEEN argued, is to concern ourselves with the representation of a representation ("images as simulacra for the text"). The TEI proponents' interest in images tends to focus on corollary materials for use in teaching, for example, photographs of the Acropolis to accompany a Greek text.

By the end of the Workshop, SPERBERG-McQUEEN confessed to having been converted to a limited extent to the view that electronic images constitute a promising alternative to microfilming; indeed, an alternative probably superior to microfilming. But he was not convinced

that electronic images constitute a serious attempt to represent text in electronic form. HOCKEY and MYLONAS also conceded that their experience at the Pierce Symposium the previous week at Georgetown University and the present conference at the Library of Congress had compelled them to reevaluate their perspective on the usefulness of text as images. Attendees could see that the text and image advocates were in constructive tension, so to say.

Three nonTEI presentations described approaches to preparing machine-readable text that are less rigorous and thus less expensive. In the case of the Papers of George Washington, Dorothy TWOHIG explained that the digital version will provide a not-quite-perfect rendering of the transcribed text--some 135,000 documents, available for research during the decades while the perfect or print version is completed. Members of the American Memory team and the staff of NAL's Text Digitization Program (see below) also outlined a middle ground concerning searchable texts. In the case of American Memory, contractors produce texts with about 99-percent accuracy that serve as "browse" or "reference" versions of written or printed originals. End users who need faithful copies or perfect renditions must refer to accompanying sets of digital facsimile images or consult copies of the originals in a nearby library or archive. American Memory staff argued that the high cost of producing 100-percent accurate copies would prevent LC from offering access to large parts of its collections.

THE MACHINE-READABLE TEXT: METHODS OF CONVERSION

Although the Workshop did not include a systematic examination of the methods for converting texts from paper (or from facsimile images) into machine-readable form, nevertheless, various speakers touched upon this matter. For example, WEIBEL reported that OCLC has experimented with a merging of multiple optical character recognition systems that will reduce errors from an unacceptable rate of 5 characters out of every l,000 to an unacceptable rate of 2 characters out of every l,000.

Pamela ANDRE presented an overview of NAL's Text Digitization Program and Judith ZIDAR discussed the technical details. ZIDAR explained how NAL purchased hardware and software capable of performing optical character recognition (OCR) and text conversion and used its own staff to convert texts. The process, ZIDAR said, required extensive editing and project staff found themselves considering alternatives, including rekeying and/or creating abstracts or summaries of texts. NAL reckoned costs at $7 per page. By way of contrast, Ricky ERWAY explained that American Memory had decided from the start to contract out conversion to external service bureaus. The criteria used to select these contractors were cost and quality of results, as opposed to methods of conversion. ERWAY noted that historical documents or books often do not lend themselves to OCR. Bound materials represent a special problem. In her experience, quality control--inspecting incoming materials, counting errors in samples--posed the most time-consuming aspect of contracting out conversion. ERWAY reckoned American Memory's costs at $4 per page, but cautioned that fewer cost-elements had been included than in NAL's figure.

OPTIONS FOR DISSEMINATION

The topic of dissemination proper emerged at various points during the Workshop. At the session devoted to national and international computer networks, LYNCH, Howard BESSER, Ronald LARSEN, and Edwin BROWNRIGG highlighted the virtues of Internet today and of the network that will evolve from Internet. Listeners could discern in these narratives a vision of an information democracy in which millions of citizens freely find and use what they need. LYNCH noted that a lack of standards inhibits disseminating multimedia on the network, a topic also discussed by BESSER. LARSEN addressed the issues of network scalability and modularity and commented upon the difficulty of anticipating the effects of growth in orders of magnitude. BROWNRIGG talked about the ability of packet radio to provide certain links in a network without the need for wiring. However, the presenters also called attention to the shortcomings and incongruities of present-day computer networks. For example: 1) Network use is growing dramatically, but much network traffic consists of personal communication (E-mail). 2) Large bodies of information are available, but a user's ability to search across their entirety is limited. 3) There are significant resources for science and technology, but few network sources provide content in the humanities. 4) Machine-readable texts are commonplace, but the capability of the system to deal with images (let alone other media formats) lags behind. A glimpse of a multimedia future for networks, however, was provided by Maria LEBRON in her overview of the Online Journal of Current Clinical Trials (OJCCT), and the process of scholarly publishing on-line.

The contrasting form of the CD-ROM disk was never systematically analyzed, but attendees could glean an impression from several of the show-and-tell presentations. The Perseus and American Memory examples demonstrated recently published disks, while the descriptions of the IBYCUS version of the Papers of George Washington and Chadwyck-Healey's Patrologia Latina Database (PLD) told of disks to come. According to Eric CALALUCA, PLD's principal focus has been on converting Jacques-Paul Migne's definitive collection of Latin texts to machine-readable form. Although everyone could share the network advocates' enthusiasm for an on-line future, the possibility of rolling up one's sleeves for a session with a CD-ROM containing both textual materials and a powerful retrieval engine made the disk seem an appealing vessel indeed. The overall discussion suggested that the transition from CD-ROM to on-line networked access may prove far slower and more difficult than has been anticipated.

WHO ARE THE USERS AND WHAT DO THEY DO?

Although concerned with the technicalities of production, the Workshop never lost sight of the purposes and uses of electronic versions of textual materials. As noted above, those interested in imaging discussed the problematical matter of digital preservation, while the TEI proponents described how machine-readable texts can be used in research. This latter topic received thorough treatment in the paper read by Avra MICHELSON. She placed the phenomenon of electronic texts within the context of broader trends in information technology and scholarly communication.

Among other things, MICHELSON described on-line conferences that represent a vigorous and important intellectual forum for certain disciplines. Internet now carries more than 700 conferences, with about 80 percent of these devoted to topics in the social sciences and the humanities. Other scholars use on-line networks for "distance learning." Meanwhile, there has been a tremendous growth in end-user computing; professors today are less likely than their predecessors to ask the campus computer center to process their data. Electronic texts are one key to these sophisticated applications, MICHELSON reported, and more and more scholars in the humanities now work in an on-line environment. Toward the end of the Workshop, Michael LESK presented a corollary to MICHELSON's talk, reporting the results of an experiment that compared the work of one group of chemistry students using traditional printed texts and two groups using electronic sources. The experiment demonstrated that in the event one does not know what to read, one needs the electronic systems; the electronic systems hold no advantage at the moment if one knows what to read, but neither do they impose a penalty.

DALY provided an anecdotal account of the revolutionizing impact of the new technology on his previous methods of research in the field of classics. His account, by extrapolation, served to illustrate in part the arguments made by MICHELSON concerning the positive effects of the sudden and radical transformation being wrought in the ways scholars work.

Susan VECCIA and Joanne FREEMAN delineated the use of electronic materials outside the university. The most interesting aspect of their

use, FREEMAN said, could be seen as a paradox: teachers in elementary and secondary schools requested access to primary source materials but, at the same time, found that "primariness" itself made these materials difficult for their students to use.

OTHER TOPICS

Marybeth PETERS reviewed copyright law in the United States and offered advice during a lively discussion of this subject. But uncertainty remains concerning the price of copyright in a digital medium, because a solution remains to be worked out concerning management and synthesis of copyrighted and out-of-copyright pieces of a database.

As moderator of the final session of the Workshop, Prosser GIFFORD directed discussion to future courses of action and the potential role of LC in advancing them. Among the recommendations that emerged were the following:

* Workshop participants should 1) begin to think about working with image material, but structure and digitize it in such a way that at a later stage it can be interpreted into text, and 2) find a common way to build text and images together so that they can be used jointly at some stage in the future, with appropriate network support, because that is how users will want to access these materials. The Library might encourage attempts to bring together people who are working on texts and images.

* A network version of American Memory should be developed or consideration should be given to making the data in it available to people interested in doing network multimedia. Given the current dearth of digital data that is appealing and unencumbered by extremely complex rights problems, developing a network version of American Memory could do much to help make network multimedia a reality.

* Concerning the thorny issue of electronic deposit, LC should initiate a catalytic process in terms of distributed responsibility, that is, bring together the distributed organizations and set up a study group to look at all the issues related to electronic deposit and see where we as a nation should move. For example, LC might attempt to persuade one major library in each state to deal with its state equivalent publisher, which might produce a cooperative project that would be equitably distributed around the country, and one in which LC would be dealing with a minimal number of publishers and minimal copyright problems. LC must also deal with the concept of on-line publishing, determining, among other things, how serials such as OJCCT might be deposited for copyright.

* Since a number of projects are planning to carry out preservation by creating digital images that will end up in on-line or near-line storage at some institution, LC might play a helpful role, at least in the near term, by accelerating how

to catalog that information into the Research Library Information Network (RLIN) and then into OCLC, so that it would be accessible. This would reduce the possibility of multiple institutions digitizing the same work.

CONCLUSION

The Workshop was valuable because it brought together partisans from various groups and provided an occasion to compare goals and methods. The more committed partisans frequently communicate with others in their groups, but less often across group boundaries. The Workshop was also valuable to attendees--including those involved with American Memory--who came less committed to particular approaches or concepts. These attendees learned a great deal, and plan to select and employ elements of imaging, text-coding, and networked distribution that suit their respective projects and purposes.

Still, reality rears its ugly head: no breakthrough has been achieved. On the imaging side, one confronts a proliferation of competing data-interchange standards and a lack of consensus on the role of digital facsimiles in preservation. In the realm of machine-readable texts, one encounters a reasonably mature standard but methodological difficulties and high costs. These latter problems, of course, represent a special impediment to the desire, as it is sometimes expressed in the popular press, "to put the [contents of the] Library of Congress on line." In the words of one participant, there was "no solution to the economic

problems--the projects that are out there are surviving, but it is going to be a lot of work to transform the information industry, and so far the investment to do that is not forthcoming" (LESK, per litteras).

*** *** *** ****** *** *** ***

PROCEEDINGS

WELCOME

++

GIFFORD * Origin of Workshop in current Librarian's desire to make LC's collections more widely available * Desiderata arising from the prospect of greater interconnectedness *

++

After welcoming participants on behalf of the Library of Congress, American Memory (AM), and the National Demonstration Lab, Prosser GIFFORD, director for scholarly programs, Library of Congress, located the origin of the Workshop on Electronic Texts in a conversation he had had considerably more than a year ago with Carl FLEISCHHAUER concerning some of the issues faced by AM. On the assumption that numerous other people were asking the same questions, the decision was made to bring together as many of these people as possible to ask the same questions together. In a deeper sense, GIFFORD said, the origin of the Workshop

lay in the desire of the current Librarian of Congress, James H. Billington, to make the collections of the Library, especially those offering unique or unusual testimony on aspects of the American experience, available to a much wider circle of users than those few people who can come to Washington to use them. This meant that the emphasis of AM, from the outset, has been on archival collections of the basic material, and on making these collections themselves available, rather than selected or heavily edited products.

From AM's emphasis followed the questions with which the Workshop began: who will use these materials, and in what form will they wish to use them. But an even larger issue deserving mention, in GIFFORD's view, was the phenomenal growth in Internet connectivity. He expressed the hope that the prospect of greater interconnectedness than ever before would lead to: 1) much more cooperative and mutually supportive endeavors; 2) development of systems of shared and distributed responsibilities to avoid duplication and to ensure accuracy and preservation of unique materials; and 3) agreement on the necessary standards and development of the appropriate directories and indices to make navigation straightforward among the varied resources that are, and increasingly will be, available. In this connection, GIFFORD requested that participants reflect from the outset upon the sorts of outcomes they thought the Workshop might have. Did those present constitute a group with sufficient common interests to propose a next step or next steps, and if so, what might those be? They would return to these questions the following afternoon.

******

++

FLEISCHHAUER * Core of Workshop concerns preparation and production of materials * Special challenge in conversion of textual materials * Quality versus quantity * Do the several groups represented share common interests? *

++

Carl FLEISCHHAUER, coordinator, American Memory, Library of Congress, emphasized that he would attempt to represent the people who perform some of the work of converting or preparing materials and that the core of the Workshop had to do with preparation and production. FLEISCHHAUER then drew a distinction between the long term, when many things would be available and connected in the ways that GIFFORD described, and the short term, in which AM not only has wrestled with the issue of what is the best course to pursue but also has faced a variety of technical challenges.

FLEISCHHAUER remarked AM's endeavors to deal with a wide range of library formats, such as motion picture collections, sound-recording collections, and pictorial collections of various sorts, especially collections of photographs. In the course of these efforts, AM kept coming back to textual materials--manuscripts or rare printed matter, bound materials, etc. Text posed the greatest conversion challenge of all. Thus, the genesis of the Workshop, which reflects the problems faced by AM. These

problems include physical problems. For example, those in the library and archive business deal with collections made up of fragile and rare manuscript items, bound materials, especially the notoriously brittle bound materials of the late nineteenth century. These are precious cultural artifacts, however, as well as interesting sources of information, and LC desires to retain and conserve them. AM needs to handle things without damaging them. Guillotining a book to run it through a sheet feeder must be avoided at all costs.

Beyond physical problems, issues pertaining to quality arose. For example, the desire to provide users with a searchable text is affected by the question of acceptable level of accuracy. One hundred percent accuracy is tremendously expensive. On the other hand, the output of optical character recognition (OCR) can be tremendously inaccurate. Although AM has attempted to find a middle ground, uncertainty persists as to whether or not it has discovered the right solution.

Questions of quality arose concerning images as well. FLEISCHHAUER contrasted the extremely high level of quality of the digital images in the Cornell Xerox Project with AM's efforts to provide a browse-quality or access-quality image, as opposed to an archival or preservation image. FLEISCHHAUER therefore welcomed the opportunity to compare notes.

FLEISCHHAUER observed in passing that conversations he had had about networks have begun to signal that for various forms of media a determination may be made that there is a browse-quality item, or a

distribution-and-access-quality item that may coexist in some systems with a higher quality archival item that would be inconvenient to send through the network because of its size. FLEISCHHAUER referred, of course, to images more than to searchable text.

As AM considered those questions, several conceptual issues arose: ought AM occasionally to reproduce materials entirely through an image set, at other times, entirely through a text set, and in some cases, a mix? There probably would be times when the historical authenticity of an artifact would require that its image be used. An image might be desirable as a recourse for users if one could not provide 100-percent accurate text. Again, AM wondered, as a practical matter, if a distinction could be drawn between rare printed matter that might exist in multiple collections--that is, in ten or fifteen libraries. In such cases, the need for perfect reproduction would be less than for unique items. Implicit in his remarks, FLEISCHHAUER conceded, was the admission that AM has been tilting strongly towards quantity and drawing back a little from perfect quality. That is, it seemed to AM that society would be better served if more things were distributed by LC--even if they were not quite perfect--than if fewer things, perfectly represented, were distributed. This was stated as a proposition to be tested, with responses to be gathered from users.

In thinking about issues related to reproduction of materials and seeing other people engaged in parallel activities, AM deemed it useful to convene a conference. Hence, the Workshop. FLEISCHHAUER thereupon surveyed the several groups represented: 1) the world of images (image

users and image makers); 2) the world of text and scholarship and, within this group, those concerned with language--FLEISCHHAUER confessed to finding delightful irony in the fact that some of the most advanced thinkers on computerized texts are those dealing with ancient Greek and Roman materials; 3) the network world; and 4) the general world of library science, which includes people interested in preservation and cataloging.

FLEISCHHAUER concluded his remarks with special thanks to the David and Lucile Packard Foundation for its support of the meeting, the American Memory group, the Office for Scholarly Programs, the National Demonstration Lab, and the Office of Special Events. He expressed the hope that David Woodley Packard might be able to attend, noting that Packard's work and the work of the foundation had sponsored a number of projects in the text area.

******

SESSION I. CONTENT IN A NEW FORM: WHO WILL USE IT AND WHAT WILL THEY DO?

++

DALY * Acknowledgements * A new Latin authors disk * Effects of the new technology on previous methods of research *

++

Serving as moderator, James DALY acknowledged the generosity of all the

presenters for giving of their time, counsel, and patience in planning the Workshop, as well as of members of the American Memory project and other Library of Congress staff, and the David and Lucile Packard Foundation and its executive director, Colburn S. Wilbur.

DALY then recounted his visit in March to the Center for Electronic Texts in the Humanities (CETH) and the Department of Classics at Rutgers University, where an old friend, Lowell Edmunds, introduced him to the department's IBYCUS scholarly personal computer, and, in particular, the new Latin CD-ROM, containing, among other things, almost all classical Latin literary texts through A.D. 200. Packard Humanities Institute (PHI), Los Altos, California, released this disk late in 1991, with a nominal triennial licensing fee.

Playing with the disk for an hour or so at Rutgers brought home to DALY at once the revolutionizing impact of the new technology on his previous methods of research. Had this disk been available two or three years earlier, DALY contended, when he was engaged in preparing a commentary on Book 10 of Virgil's Aeneid for Cambridge University Press, he would not have required a forty-eight-square-foot table on which to spread the numerous, most frequently consulted items, including some ten or twelve concordances to key Latin authors, an almost equal number of lexica to authors who lacked concordances, and where either lexica or concordances were lacking, numerous editions of authors antedating and postdating Virgil.

Nor, when checking each of the average six to seven words contained in

the Virgilian hexameter for its usage elsewhere in Virgil's works or other Latin authors, would DALY have had to maintain the laborious mechanical process of flipping through these concordances, lexica, and editions each time. Nor would he have had to frequent as often the Milton S. Eisenhower Library at the Johns Hopkins University to consult the Thesaurus Linguae Latinae. Instead of devoting countless hours, or the bulk of his research time, to gathering data concerning Virgil's use of words, DALY--now freed by PHI's Latin authors disk from the tyrannical, yet in some ways paradoxically happy scholarly drudgery--would have been able to devote that same bulk of time to analyzing and interpreting Virgilian verbal usage.

Citing Theodore Brunner, Gregory Crane, Elli MYLONAS, and Avra MICHELSON, DALY argued that this reversal in his style of work, made possible by the new technology, would perhaps have resulted in better, more productive research. Indeed, even in the course of his browsing the Latin authors disk at Rutgers, its powerful search, retrieval, and highlighting capabilities suggested to him several new avenues of research into Virgil's use of sound effects. This anecdotal account, DALY maintained, may serve to illustrate in part the sudden and radical transformation being wrought in the ways scholars work.

******

++

MICHELSON * Elements related to scholarship and technology * Electronic

texts within the context of broader trends within information technology and scholarly communication * Evaluation of the prospects for the use of electronic texts * Relationship of electronic texts to processes of scholarly communication in humanities research * New exchange formats created by scholars * Projects initiated to increase scholarly access to converted text * Trend toward making electronic resources available through research and education networks * Changes taking place in scholarly communication among humanities scholars * Network-mediated scholarship transforming traditional scholarly practices * Key information technology trends affecting the conduct of scholarly communication over the next decade * The trend toward end-user computing * The trend toward greater connectivity * Effects of these trends * Key transformations taking place * Summary of principal arguments *

++

Avra MICHELSON, Archival Research and Evaluation Staff, National Archives and Records Administration (NARA), argued that establishing who will use electronic texts and what they will use them for involves a consideration of both information technology and scholarship trends. This consideration includes several elements related to scholarship and technology: 1) the key trends in information technology that are most relevant to scholarship; 2) the key trends in the use of currently available technology by scholars in the nonscientific community; and 3) the relationship between these two very distinct but interrelated trends. The investment in understanding this relationship being made by information providers, technologists, and public policy developers, as well as by scholars themselves, seems to be pervasive and growing,

MICHELSON contended. She drew on collaborative work with Jeff Rothenberg on the scholarly use of technology.

MICHELSON sought to place the phenomenon of electronic texts within the context of broader trends within information technology and scholarly communication. She argued that electronic texts are of most use to researchers to the extent that the researchers' working context (i.e., their relevant bibliographic sources, collegial feedback, analytic tools, notes, drafts, etc.), along with their field's primary and secondary sources, also is accessible in electronic form and can be integrated in ways that are unique to the on-line environment.

Evaluation of the prospects for the use of electronic texts includes two elements: 1) an examination of the ways in which researchers currently are using electronic texts along with other electronic resources, and 2) an analysis of key information technology trends that are affecting the long-term conduct of scholarly communication. MICHELSON limited her discussion of the use of electronic texts to the practices of humanists and noted that the scientific community was outside the panel's overview.

MICHELSON examined the nature of the current relationship of electronic texts in particular, and electronic resources in general, to what she maintained were, essentially, five processes of scholarly communication in humanities research. Researchers 1) identify sources, 2) communicate with their colleagues, 3) interpret and analyze data, 4) disseminate their research findings, and 5) prepare curricula to instruct the next

generation of scholars and students. This examination would produce a clearer understanding of the synergy among these five processes that fuels the tendency of the use of electronic resources for one process to stimulate its use for other processes of scholarly communication.

For the first process of scholarly communication, the identification of sources, MICHELSON remarked the opportunity scholars now enjoy to supplement traditional word-of-mouth searches for sources among their colleagues with new forms of electronic searching. So, for example, instead of having to visit the library, researchers are able to explore descriptions of holdings in their offices. Furthermore, if their own institutions' holdings prove insufficient, scholars can access more than 200 major American library catalogues over Internet, including the universities of California, Michigan, Pennsylvania, and Wisconsin. Direct access to the bibliographic databases offers intellectual empowerment to scholars by presenting a comprehensive means of browsing through libraries from their homes and offices at their convenience.

The second process of communication involves communication among scholars. Beyond the most common methods of communication, scholars are using E-mail and a variety of new electronic communications formats derived from it for further academic interchange. E-mail exchanges are growing at an astonishing rate, reportedly 15 percent a month. They currently constitute approximately half the traffic on research and education networks. Moreover, the global spread of E-mail has been so rapid that it is now possible for American scholars to use it to communicate with colleagues in close to 140 other countries.

Other new exchange formats created by scholars and operating on Internet include more than 700 conferences, with about 80 percent of these devoted to topics in the social sciences and humanities. The rate of growth of these scholarly electronic conferences also is astonishing. From l990 to l991, 200 new conferences were identified on Internet. From October 1991 to June 1992, an additional 150 conferences in the social sciences and humanities were added to this directory of listings. Scholars have established conferences in virtually every field, within every different discipline. For example, there are currently close to 600 active social science and humanities conferences on topics such as art and architecture, ethnomusicology, folklore, Japanese culture, medical education, and gifted and talented education. The appeal to scholars of communicating through these conferences is that, unlike any other medium, electronic conferences today provide a forum for global communication with peers at the front end of the research process.

Interpretation and analysis of sources constitutes the third process of scholarly communication that MICHELSON discussed in terms of texts and textual resources. The methods used to analyze sources fall somewhere on a continuum from quantitative analysis to qualitative analysis. Typically, evidence is culled and evaluated using methods drawn from both ends of this continuum. At one end, quantitative analysis involves the use of mathematical processes such as a count of frequencies and distributions of occurrences or, on a higher level, regression analysis. At the other end of the continuum, qualitative analysis typically involves nonmathematical processes oriented toward language

interpretation or the building of theory. Aspects of this work involve the processing--either manual or computational--of large and sometimes massive amounts of textual sources, although the use of nontextual sources as evidence, such as photographs, sound recordings, film footage, and artifacts, is significant as well.

Scholars have discovered that many of the methods of interpretation and analysis that are related to both quantitative and qualitative methods are processes that can be performed by computers. For example, computers can count. They can count brush strokes used in a Rembrandt painting or perform regression analysis for understanding cause and effect. By means of advanced technologies, computers can recognize patterns, analyze text, and model concepts. Furthermore, computers can complete these processes faster with more sources and with greater precision than scholars who must rely on manual interpretation of data. But if scholars are to use computers for these processes, source materials must be in a form amenable to computer-assisted analysis. For this reason many scholars, once they have identified the sources that are key to their research, are converting them to machine-readable form. Thus, a representative example of the numerous textual conversion projects organized by scholars around the world in recent years to support computational text analysis is the TLG, the Thesaurus Linguae Graecae. This project is devoted to converting the extant ancient texts of classical Greece. (Editor's note: according to the TLG Newsletter of May l992, TLG was in use in thirty-two different countries. This figure updates MICHELSON's previous count by one.)

The scholars performing these conversions have been asked to recognize

that the electronic sources they are converting for one use possess value for other research purposes as well. As a result, during the past few years, humanities scholars have initiated a number of projects to increase scholarly access to converted text. So, for example, the Text Encoding Initiative (TEI), about which more is said later in the program, was established as an effort by scholars to determine standard elements and methods for encoding machine-readable text for electronic exchange. In a second effort to facilitate the sharing of converted text, scholars have created a new institution, the Center for Electronic Texts in the Humanities (CETH). The center estimates that there are 8,000 series of source texts in the humanities that have been converted to machine-readable form worldwide. CETH is undertaking an international search for converted text in the humanities, compiling it into an electronic library, and preparing bibliographic descriptions of the sources for the Research Libraries Information Network's (RLIN) machine-readable data file. The library profession has begun to initiate large conversion projects as well, such as American Memory.

While scholars have been making converted text available to one another, typically on disk or on CD-ROM, the clear trend is toward making these resources available through research and education networks. Thus, the American and French Research on the Treasury of the French Language (ARTFL) and the Dante Project are already available on Internet. MICHELSON summarized this section on interpretation and analysis by noting that: 1) increasing numbers of humanities scholars in the library community are recognizing the importance to the advancement of scholarship of retrospective conversion of source materials in the arts

and humanities; and 2) there is a growing realization that making the sources available on research and education networks maximizes their usefulness for the analysis performed by humanities scholars.

The fourth process of scholarly communication is dissemination of research findings, that is, publication. Scholars are using existing research and education networks to engineer a new type of publication: scholarly-controlled journals that are electronically produced and disseminated. Although such journals are still emerging as a communication format, their number has grown, from approximately twelve to thirty-six during the past year (July 1991 to June 1992). Most of these electronic scholarly journals are devoted to topics in the humanities. As with network conferences, scholarly enthusiasm for these electronic journals stems from the medium's unique ability to advance scholarship in a way that no other medium can do by supporting global feedback and interchange, practically in real time, early in the research process. Beyond scholarly journals, MICHELSON remarked the delivery of commercial full-text products, such as articles in professional journals, newsletters, magazines, wire services, and reference sources. These are being delivered via on-line local library catalogues, especially through CD-ROMs. Furthermore, according to MICHELSON, there is general optimism that the copyright and fees issues impeding the delivery of full text on existing research and education networks soon will be resolved.

The final process of scholarly communication is curriculum development and instruction, and this involves the use of computer information technologies in two areas. The first is the development of

computer-oriented instructional tools, which includes simulations, multimedia applications, and computer tools that are used to assist in the analysis of sources in the classroom, etc. The Perseus Project, a database that provides a multimedia curriculum on classical Greek civilization, is a good example of the way in which entire curricula are being recast using information technologies. It is anticipated that the current difficulty in exchanging electronically computer-based instructional software, which in turn makes it difficult for one scholar to build upon the work of others, will be resolved before too long. Stand-alone curricular applications that involve electronic text will be sharable through networks, reinforcing their significance as intellectual products as well as instructional tools.

The second aspect of electronic learning involves the use of research and education networks for distance education programs. Such programs interactively link teachers with students in geographically scattered locations and rely on the availability of electronic instructional resources. Distance education programs are gaining wide appeal among state departments of education because of their demonstrated capacity to bring advanced specialized course work and an array of experts to many classrooms. A recent report found that at least 32 states operated at least one statewide network for education in 1991, with networks under development in many of the remaining states.

MICHELSON summarized this section by noting two striking changes taking place in scholarly communication among humanities scholars. First is the extent to which electronic text in particular, and electronic resources

in general, are being infused into each of the five processes described above. As mentioned earlier, there is a certain synergy at work here. The use of electronic resources for one process tends to stimulate its use for other processes, because the chief course of movement is toward a comprehensive on-line working context for humanities scholars that includes on-line availability of key bibliographies, scholarly feedback, sources, analytical tools, and publications. MICHELSON noted further that the movement toward a comprehensive on-line working context for humanities scholars is not new. In fact, it has been underway for more than forty years in the humanities, since Father Roberto Busa began developing an electronic concordance of the works of Saint Thomas Aquinas in 1949. What we are witnessing today, MICHELSON contended, is not the beginning of this on-line transition but, for at least some humanities scholars, the turning point in the transition from a print to an electronic working context. Coinciding with the on-line transition, the second striking change is the extent to which research and education networks are becoming the new medium of scholarly communication. The existing Internet and the pending National Education and Research Network (NREN) represent the new meeting ground where scholars are going for bibliographic information, scholarly dialogue and feedback, the most current publications in their field, and high-level educational offerings. Traditional scholarly practices are undergoing tremendous transformations as a result of the emergence and growing prominence of what is called network-mediated scholarship.

MICHELSON next turned to the second element of the framework she proposed at the outset of her talk for evaluating the prospects for electronic

text, namely the key information technology trends affecting the conduct of scholarly communication over the next decade: 1) end-user computing and 2) connectivity.

End-user computing means that the person touching the keyboard, or performing computations, is the same as the person who initiates or consumes the computation. The emergence of personal computers, along with a host of other forces, such as ubiquitous computing, advances in interface design, and the on-line transition, is prompting the consumers of computation to do their own computing, and is thus rendering obsolete the traditional distinction between end users and ultimate users.

The trend toward end-user computing is significant to consideration of the prospects for electronic texts because it means that researchers are becoming more adept at doing their own computations and, thus, more competent in the use of electronic media. By avoiding programmer intermediaries, computation is becoming central to the researcher's thought process. This direct involvement in computing is changing the researcher's perspective on the nature of research itself, that is, the kinds of questions that can be posed, the analytical methodologies that can be used, the types and amount of sources that are appropriate for analyses, and the form in which findings are presented. The trend toward end-user computing means that, increasingly, electronic media and computation are being infused into all processes of humanities scholarship, inspiring remarkable transformations in scholarly communication.

The trend toward greater connectivity suggests that researchers are using computation increasingly in network environments. Connectivity is important to scholarship because it erases the distance that separates students from teachers and scholars from their colleagues, while allowing users to access remote databases, share information in many different media, connect to their working context wherever they are, and collaborate in all phases of research.

The combination of the trend toward end-user computing and the trend toward connectivity suggests that the scholarly use of electronic resources, already evident among some researchers, will soon become an established feature of scholarship. The effects of these trends, along with ongoing changes in scholarly practices, point to a future in which humanities researchers will use computation and electronic communication to help them formulate ideas, access sources, perform research, collaborate with colleagues, seek peer review, publish and disseminate results, and engage in many other professional and educational activities.

In summary, MICHELSON emphasized four points: 1) A portion of humanities scholars already consider electronic texts the preferred format for analysis and dissemination. 2) Scholars are using these electronic texts, in conjunction with other electronic resources, in all the processes of scholarly communication. 3) The humanities scholars' working context is in the process of changing from print technology to electronic technology, in many ways mirroring transformations that have occurred or are occurring within the scientific community. 4) These

changes are occurring in conjunction with the development of a new communication medium: research and education networks that are characterized by their capacity to advance scholarship in a wholly unique way.

MICHELSON also reiterated her three principal arguments: l) Electronic texts are best understood in terms of the relationship to other electronic resources and the growing prominence of network-mediated scholarship. 2) The prospects for electronic texts lie in their capacity to be integrated into the on-line network of electronic resources that comprise the new working context for scholars. 3) Retrospective conversion of portions of the scholarly record should be a key strategy as information providers respond to changes in scholarly communication practices.

******

++

VECCIA * AM's evaluation project and public users of electronic resources * AM and its design * Site selection and evaluating the Macintosh implementation of AM * Characteristics of the six public libraries selected * Characteristics of AM's users in these libraries * Principal ways AM is being used *

++

Susan VECCIA, team leader, and Joanne FREEMAN, associate coordinator, American Memory, Library of Congress, gave a joint presentation. First,

by way of introduction, VECCIA explained her and FREEMAN's roles in American Memory (AM). Serving principally as an observer, VECCIA has assisted with the evaluation project of AM, placing AM collections in a variety of different sites around the country and helping to organize and implement that project. FREEMAN has been an associate coordinator of AM and has been involved principally with the interpretative materials, preparing some of the electronic exhibits and printed historical information that accompanies AM and that is requested by users. VECCIA and FREEMAN shared anecdotal observations concerning AM with public users of electronic resources. Notwithstanding a fairly structured evaluation in progress, both VECCIA and FREEMAN chose not to report on specifics in terms of numbers, etc., because they felt it was too early in the evaluation project to do so.

AM is an electronic archive of primary source materials from the Library of Congress, selected collections representing a variety of formats--photographs, graphic arts, recorded sound, motion pictures, broadsides, and soon, pamphlets and books. In terms of the design of this system, the interpretative exhibits have been kept separate from the primary resources, with good reason. Accompanying this collection are printed documentation and user guides, as well as guides that FREEMAN prepared for teachers so that they may begin using the content of the system at once.

VECCIA described the evaluation project before talking about the public users of AM, limiting her remarks to public libraries, because FREEMAN would talk more specifically about schools from kindergarten to twelfth grade (K-12). Having started in spring 1991, the evaluation currently

involves testing of the Macintosh implementation of AM. Since the primary goal of this evaluation is to determine the most appropriate audience or audiences for AM, very different sites were selected. This makes evaluation difficult because of the varying degrees of technology literacy among the sites. AM is situated in forty-four locations, of which six are public libraries and sixteen are schools. Represented among the schools are elementary, junior high, and high schools. District offices also are involved in the evaluation, which will conclude in summer 1993.

VECCIA focused the remainder of her talk on the six public libraries, one of which doubles as a state library. They represent a range of geographic areas and a range of demographic characteristics. For example, three are located in urban settings, two in rural settings, and one in a suburban setting. A range of technical expertise is to be found among these facilities as well. For example, one is an "Apple library of the future," while two others are rural one-room libraries--in one, AM sits at the front desk next to a tractor manual.

All public libraries have been extremely enthusiastic, supportive, and appreciative of the work that AM has been doing. VECCIA characterized various users: Most users in public libraries describe themselves as general readers; of the students who use AM in the public libraries, those in fourth grade and above seem most interested. Public libraries in rural sites tend to attract retired people, who have been highly receptive to AM. Users tend to fall into two additional categories: people interested in the content and historical connotations of these

primary resources, and those fascinated by the technology. The format receiving the most comments has been motion pictures. The adult users in public libraries are more comfortable with IBM computers, whereas young people seem comfortable with either IBM or Macintosh, although most of them seem to come from a Macintosh background. This same tendency is found in the schools.

What kinds of things do users do with AM? In a public library there are two main goals or ways that AM is being used: as an individual learning tool, and as a leisure activity. Adult learning was one area that VECCIA would highlight as a possible application for a tool such as AM. She described a patron of a rural public library who comes in every day on his lunch hour and literally reads AM, methodically going through the collection image by image. At the end of his hour he makes an electronic bookmark, puts it in his pocket, and returns to work. The next day he comes in and resumes where he left off. Interestingly, this man had never been in the library before he used AM. In another small, rural library, the coordinator reports that AM is a popular activity for some of the older, retired people in the community, who ordinarily would not use "those things,"--computers. Another example of adult learning in public libraries is book groups, one of which, in particular, is using AM as part of its reading on industrialization, integration, and urbanization in the early 1900s.

One library reports that a family is using AM to help educate their children. In another instance, individuals from a local museum came in to use AM to prepare an exhibit on toys of the past. These two examples

emphasize the mission of the public library as a cultural institution, reaching out to people who do not have the same resources available to those who live in a metropolitan area or have access to a major library. One rural library reports that junior high school students in large numbers came in one afternoon to use AM for entertainment. A number of public libraries reported great interest among postcard collectors in the Detroit collection, which was essentially a collection of images used on postcards around the turn of the century. Train buffs are similarly interested because that was a time of great interest in railroading. People, it was found, relate to things that they know of firsthand. For example, in both rural public libraries where AM was made available, observers reported that the older people with personal remembrances of the turn of the century were gravitating to the Detroit collection. These examples served to underscore MICHELSON's observation re the integration of electronic tools and ideas--that people learn best when the material relates to something they know.

VECCIA made the final point that in many cases AM serves as a public-relations tool for the public libraries that are testing it. In one case, AM is being used as a vehicle to secure additional funding for the library. In another case, AM has served as an inspiration to the staff of a major local public library in the South to think about ways to make its own collection of photographs more accessible to the public.

******

++

FREEMAN * AM and archival electronic resources in a school environment *
Questions concerning context * Questions concerning the electronic format
itself * Computer anxiety * Access and availability of the system *
Hardware * Strengths gained through the use of archival resources in
schools *

++

Reiterating an observation made by VECCIA, that AM is an archival resource made up of primary materials with very little interpretation, FREEMAN stated that the project has attempted to bridge the gap between these bare primary materials and a school environment, and in that cause has created guided introductions to AM collections. Loud demand from the educational community, chiefly from teachers working with the upper grades of elementary school through high school, greeted the announcement that AM would be tested around the country.

FREEMAN reported not only on what was learned about AM in a school environment, but also on several universal questions that were raised concerning archival electronic resources in schools. She discussed several strengths of this type of material in a school environment as opposed to a highly structured resource that offers a limited number of paths to follow.

FREEMAN first raised several questions about using AM in a school environment. There is often some difficulty in developing a sense of

what the system contains. Many students sit down at a computer resource and assume that, because AM comes from the Library of Congress, all of American history is now at their fingertips. As a result of that sort of mistaken judgment, some students are known to conclude that AM contains nothing of use to them when they look for one or two things and do not find them. It is difficult to discover that middle ground where one has a sense of what the system contains. Some students grope toward the idea of an archive, a new idea to them, since they have not previously experienced what it means to have access to a vast body of somewhat random information.

Other questions raised by FREEMAN concerned the electronic format itself. For instance, in a school environment it is often difficult both for teachers and students to gain a sense of what it is they are viewing. They understand that it is a visual image, but they do not necessarily know that it is a postcard from the turn of the century, a panoramic photograph, or even machine-readable text of an eighteenth-century broadside, a twentieth-century printed book, or a nineteenth-century diary. That distinction is often difficult for people in a school environment to grasp. Because of that, it occasionally becomes difficult to draw conclusions from what one is viewing.

FREEMAN also noted the obvious fear of the computer, which constitutes a difficulty in using an electronic resource. Though students in general did not suffer from this anxiety, several older students feared that they were computer-illiterate, an assumption that became self-fulfilling when they searched for something but failed to find it. FREEMAN said she

believed that some teachers also fear computer resources, because they believe they lack complete control. FREEMAN related the example of teachers shooing away students because it was not their time to use the system. This was a case in which the situation had to be extremely structured so that the teachers would not feel that they had lost their grasp on what the system contained.

A final question raised by FREEMAN concerned access and availability of the system. She noted the occasional existence of a gap in communication between school librarians and teachers. Often AM sits in a school library and the librarian is the person responsible for monitoring the system. Teachers do not always take into their world new library resources about which the librarian is excited. Indeed, at the sites where AM had been used most effectively within a library, the librarian was required to go to specific teachers and instruct them in its use. As a result, several AM sites will have in-service sessions over a summer, in the hope that perhaps, with a more individualized link, teachers will be more likely to use the resource.

A related issue in the school context concerned the number of workstations available at any one location. Centralization of equipment at the district level, with teachers invited to download things and walk away with them, proved unsuccessful because the hours these offices were open were also school hours.

Another issue was hardware. As VECCIA observed, a range of sites exists,

some technologically advanced and others essentially acquiring their first computer for the primary purpose of using it in conjunction with AM's testing. Users at technologically sophisticated sites want even more sophisticated hardware, so that they can perform even more sophisticated tasks with the materials in AM. But once they acquire a newer piece of hardware, they must learn how to use that also; at an unsophisticated site it takes an extremely long time simply to become accustomed to the computer, not to mention the program offered with the computer. All of these small issues raise one large question, namely, are systems like AM truly rewarding in a school environment, or do they simply act as innovative toys that do little more than spark interest?

FREEMAN contended that the evaluation project has revealed several strengths that were gained through the use of archival resources in schools, including:

* Psychic rewards from using AM as a vast, rich database, with teachers assigning various projects to students--oral presentations, written reports, a documentary, a turn-of-the-century newspaper--projects that start with the materials in AM but are completed using other resources; AM thus is used as a research tool in conjunction with other electronic resources, as well as with books and items in the library where the system is set up.

* Students are acquiring computer literacy in a humanities context.

* This sort of system is overcoming the isolation between disciplines

that often exists in schools. For example, many English teachers are requiring their students to write papers on historical topics represented in AM. Numerous teachers have reported that their students are learning critical thinking skills using the system.

* On a broader level, AM is introducing primary materials, not only to students but also to teachers, in an environment where often simply none exist--an exciting thing for the students because it helps them learn to conduct research, to interpret, and to draw their own conclusions. In learning to conduct research and what it means, students are motivated to seek knowledge. That relates to another positive outcome--a high level of personal involvement of students with the materials in this system and greater motivation to conduct their own research and draw their own conclusions.

* Perhaps the most ironic strength of these kinds of archival electronic resources is that many of the teachers AM interviewed were desperate, it is no exaggeration to say, not only for primary materials but for unstructured primary materials. These would, they thought, foster personally motivated research, exploration, and excitement in their students. Indeed, these materials have done just that. Ironically, however, this lack of structure produces some of the confusion to which the newness of these kinds of resources may also contribute. The key to effective use of archival products in a school environment is a clear, effective introduction to the system and to what it contains.

******

++

DISCUSSION * Nothing known, quantitatively, about the number of humanities scholars who must see the original versus those who would settle for an edited transcript, or about the ways in which humanities scholars are using information technology * Firm conclusions concerning the manner and extent of the use of supporting materials in print provided by AM to await completion of evaluative study * A listener's reflections on additional applications of electronic texts * Role of electronic resources in teaching elementary research skills to students *

++

During the discussion that followed the presentations by MICHELSON, VECCIA, and FREEMAN, additional points emerged.

LESK asked if MICHELSON could give any quantitative estimate of the number of humanities scholars who must see or want to see the original, or the best possible version of the material, versus those who typically would settle for an edited transcript. While unable to provide a figure, she offered her impressions as an archivist who has done some reference work and has discussed this issue with other archivists who perform reference, that those who use archives and those who use primary sources for what would be considered very high-level scholarly research, as opposed to, say, undergraduate papers, were few in number, especially

given the public interest in using primary sources to conduct genealogical or avocational research and the kind of professional research done by people in private industry or the federal government. More important in MICHELSON's view was that, quantitatively, nothing is known about the ways in which, for example, humanities scholars are using information technology. No studies exist to offer guidance in creating strategies. The most recent study was conducted in 1985 by the American Council of Learned Societies (ACLS), and what it showed was that 50 percent of humanities scholars at that time were using computers. That constitutes the extent of our knowledge.

Concerning AM's strategy for orienting people toward the scope of electronic resources, FREEMAN could offer no hard conclusions at this point, because she and her colleagues were still waiting to see, particularly in the schools, what has been made of their efforts. Within the system, however, AM has provided what are called electronic exhibits--such as introductions to time periods and materials--and these are intended to offer a student user a sense of what a broadside is and what it might tell her or him. But FREEMAN conceded that the project staff would have to talk with students next year, after teachers have had a summer to use the materials, and attempt to discover what the students were learning from the materials. In addition, FREEMAN described supporting materials in print provided by AM at the request of local teachers during a meeting held at LC. These included time lines, bibliographies, and other materials that could be reproduced on a photocopier in a classroom. Teachers could walk away with and use these, and in this way gain a better understanding of the contents. But again,

reaching firm conclusions concerning the manner and extent of their use would have to wait until next year.

As to the changes she saw occurring at the National Archives and Records Administration (NARA) as a result of the increasing emphasis on technology in scholarly research, MICHELSON stated that NARA at this point was absorbing the report by her and Jeff Rothenberg addressing strategies for the archival profession in general, although not for the National Archives specifically. NARA is just beginning to establish its role and what it can do. In terms of changes and initiatives that NARA can take, no clear response could be given at this time.

GREENFIELD remarked two trends mentioned in the session. Reflecting on DALY's opening comments on how he could have used a Latin collection of text in an electronic form, he said that at first he thought most scholars would be unwilling to do that. But as he thought of that in terms of the original meaning of research--that is, having already mastered these texts, researching them for critical and comparative purposes--for the first time, the electronic format made a lot of sense. GREENFIELD could envision growing numbers of scholars learning the new technologies for that very aspect of their scholarship and for convenience's sake.

Listening to VECCIA and FREEMAN, GREENFIELD thought of an additional application of electronic texts. He realized that AM could be used as a guide to lead someone to original sources. Students cannot be expected to have mastered these sources, things they have never known about

before. Thus, AM is leading them, in theory, to a vast body of information and giving them a superficial overview of it, enabling them to select parts of it. GREENFIELD asked if any evidence exists that this resource will indeed teach the new user, the K-12 students, how to do research. Scholars already know how to do research and are applying these new tools. But he wondered why students would go beyond picking out things that were most exciting to them.

FREEMAN conceded the correctness of GREENFIELD's observation as applied to a school environment. The risk is that a student would sit down at a system, play with it, find some things of interest, and then walk away. But in the relatively controlled situation of a school library, much will depend on the instructions a teacher or a librarian gives a student. She viewed the situation not as one of fine-tuning research skills but of involving students at a personal level in understanding and researching things. Given the guidance one can receive at school, it then becomes possible to teach elementary research skills to students, which in fact one particular librarian said she was teaching her fifth graders. FREEMAN concluded that introducing the idea of following one's own path of inquiry, which is essentially what research entails, involves more than teaching specific skills. To these comments VECCIA added the observation that the individual teacher and the use of a creative resource, rather than AM itself, seemed to make the key difference. Some schools and some teachers are making excellent use of the nature of critical thinking and teaching skills, she said.

Concurring with these remarks, DALY closed the session with the thought that

the more that producers produced for teachers and for scholars to use with their students, the more successful their electronic products would prove.

******

SESSION II. SHOW AND TELL

Jacqueline HESS, director, National Demonstration Laboratory, served as moderator of the "show-and-tell" session. She noted that a question-and-answer period would follow each presentation.

++

MYLONAS * Overview and content of Perseus * Perseus' primary materials exist in a system-independent, archival form * A concession * Textual aspects of Perseus * Tools to use with the Greek text * Prepared indices and full-text searches in Perseus * English-Greek word search leads to close study of words and concepts * Navigating Perseus by tracing down indices * Using the iconography to perform research *

++

Elli MYLONAS, managing editor, Perseus Project, Harvard University, first gave an overview of Perseus, a large, collaborative effort based at Harvard University but with contributors and collaborators located at numerous universities and colleges in the United States (e.g., Bowdoin, Maryland, Pomona, Chicago, Virginia). Funded primarily by the

Annenberg/CPB Project, with additional funding from Apple, Harvard, and the Packard Humanities Institute, among others, Perseus is a multimedia, hypertextual database for teaching and research on classical Greek civilization, which was released in February 1992 in version 1.0 and distributed by Yale University Press.

Consisting entirely of primary materials, Perseus includes ancient Greek texts and translations of those texts; catalog entries--that is, museum catalog entries, not library catalog entries--on vases, sites, coins, sculpture, and archaeological objects; maps; and a dictionary, among other sources. The number of objects and the objects for which catalog entries exist are accompanied by thousands of color images, which constitute a major feature of the database. Perseus contains approximately 30 megabytes of text, an amount that will double in subsequent versions. In addition to these primary materials, the Perseus Project has been building tools for using them, making access and navigation easier, the goal being to build part of the electronic environment discussed earlier in the morning in which students or scholars can work with their sources.

The demonstration of Perseus will show only a fraction of the real work that has gone into it, because the project had to face the dilemma of what to enter when putting something into machine-readable form: should one aim for very high quality or make concessions in order to get the material in? Since Perseus decided to opt for very high quality, all of its primary materials exist in a system-independent--insofar as it is possible to be system-independent--archival form. Deciding what that

archival form would be and attaining it required much work and thought. For example, all the texts are marked up in SGML, which will be made compatible with the guidelines of the Text Encoding Initiative (TEI) when they are issued.

Drawings are postscript files, not meeting international standards, but at least designed to go across platforms. Images, or rather the real archival forms, consist of the best available slides, which are being digitized. Much of the catalog material exists in database form--a form that the average user could use, manipulate, and display on a personal computer, but only at great cost. Thus, this is where the concession comes in: All of this rich, well-marked-up information is stripped of much of its content; the images are converted into bit-maps and the text into small formatted chunks. All this information can then be imported into HyperCard and run on a mid-range Macintosh, which is what Perseus users have. This fact has made it possible for Perseus to attain wide use fairly rapidly. Without those archival forms the HyperCard version being demonstrated could not be made easily, and the project could not have the potential to move to other forms and machines and software as they appear, none of which information is in Perseus on the CD.

Of the numerous multimedia aspects of Perseus, MYLONAS focused on the textual. Part of what makes Perseus such a pleasure to use, MYLONAS said, is this effort at seamless integration and the ability to move around both visual and textual material. Perseus also made the decision not to attempt to interpret its material any more than one interprets by selecting. But, MYLONAS emphasized, Perseus is not courseware: No

syllabus exists. There is no effort to define how one teaches a topic using Perseus, although the project may eventually collect papers by people who have used it to teach. Rather, Perseus aims to provide primary material in a kind of electronic library, an electronic sandbox, so to say, in which students and scholars who are working on this material can explore by themselves. With that, MYLONAS demonstrated Perseus, beginning with the Perseus gateway, the first thing one sees upon opening Perseus--an effort in part to solve the contextualizing problem--which tells the user what the system contains.

MYLONAS demonstrated only a very small portion, beginning with primary texts and running off the CD-ROM. Having selected Aeschylus' Prometheus Bound, which was viewable in Greek and English pretty much in the same segments together, MYLONAS demonstrated tools to use with the Greek text, something not possible with a book: looking up the dictionary entry form of an unfamiliar word in Greek after subjecting it to Perseus' morphological analysis for all the texts. After finding out about a word, a user may then decide to see if it is used anywhere else in Greek. Because vast amounts of indexing support all of the primary material, one can find out where else all forms of a particular Greek word appear--often not a trivial matter because Greek is highly inflected. Further, since the story of Prometheus has to do with the origins of sacrifice, a user may wish to study and explore sacrifice in Greek literature; by typing sacrifice into a small window, a user goes to the English-Greek word list--something one cannot do without the computer (Perseus has indexed the definitions of its dictionary)--the string sacrifice appears in the definitions of these sixty-five words. One may then find out

where any of those words is used in the work(s) of a particular author. The English definitions are not lemmatized.

All of the indices driving this kind of usage were originally devised for speed, MYLONAS observed; in other words, all that kind of information--all forms of all words, where they exist, the dictionary form they belong to--were collected into databases, which will expedite searching. Then it was discovered that one can do things searching in these databases that could not be done searching in the full texts. Thus, although there are full-text searches in Perseus, much of the work is done behind the scenes, using prepared indices. Re the indexing that is done behind the scenes, MYLONAS pointed out that without the SGML forms of the text, it could not be done effectively. Much of this indexing is based on the structures that are made explicit by the SGML tagging.

It was found that one of the things many of Perseus' non-Greek-reading users do is start from the dictionary and then move into the close study of words and concepts via this kind of English-Greek word search, by which means they might select a concept. This exercise has been assigned to students in core courses at Harvard--to study a concept by looking for the English word in the dictionary, finding the Greek words, and then finding the words in the Greek but, of course, reading across in the English. That tells them a great deal about what a translation means as well.

Should one also wish to see images that have to do with sacrifice, that person would go to the object key word search, which allows one to

perform a similar kind of index retrieval on the database of archaeological objects. Without words, pictures are useless; Perseus has not reached the point where it can do much with images that are not cataloged. Thus, although it is possible in Perseus with text and images to navigate by knowing where one wants to end up--for example, a red-figure vase from the Boston Museum of Fine Arts--one can perform this kind of navigation very easily by tracing down indices. MYLONAS illustrated several generic scenes of sacrifice on vases. The features demonstrated derived from Perseus 1.0; version 2.0 will implement even better means of retrieval.

MYLONAS closed by looking at one of the pictures and noting again that one can do a great deal of research using the iconography as well as the texts. For instance, students in a core course at Harvard this year were highly interested in Greek concepts of foreigners and representations of non-Greeks. So they performed a great deal of research, both with texts (e.g., Herodotus) and with iconography on vases and coins, on how the Greeks portrayed non-Greeks. At the same time, art historians who study iconography were also interested, and were able to use this material.

******

++

DISCUSSION * Indexing and searchability of all English words in Perseus * Several features of Perseus 1.0 * Several levels of customization possible * Perseus used for general education * Perseus' effects on

education * Contextual information in Perseus * Main challenge and emphasis of Perseus *

++

Several points emerged in the discussion that followed MYLONAS's presentation.

Although MYLONAS had not demonstrated Perseus' ability to cross-search documents, she confirmed that all English words in Perseus are indexed and can be searched. So, for example, sacrifice could have been searched in all texts, the historical essay, and all the catalogue entries with their descriptions--in short, in all of Perseus.

Boolean logic is not in Perseus 1.0 but will be added to the next version, although an effort is being made not to restrict Perseus to a database in which one just performs searching, Boolean or otherwise. It is possible to move laterally through the documents by selecting a word one is interested in and selecting an area of information one is interested in and trying to look that word up in that area.

Since Perseus was developed in HyperCard, several levels of customization are possible. Simple authoring tools exist that allow one to create annotated paths through the information, which are useful for note-taking and for guided tours for teaching purposes and for expository writing. With a little more ingenuity it is possible to begin to add or substitute material in Perseus.

Perseus has not been used so much for classics education as for general education, where it seemed to have an impact on the students in the core course at Harvard (a general required course that students must take in certain areas). Students were able to use primary material much more.

The Perseus Project has an evaluation team at the University of Maryland that has been documenting Perseus' effects on education. Perseus is very popular, and anecdotal evidence indicates that it is having an effect at places other than Harvard, for example, test sites at Ball State University, Drury College, and numerous small places where opportunities to use vast amounts of primary data may not exist. One documented effect is that archaeological, anthropological, and philological research is being done by the same person instead of by three different people.

The contextual information in Perseus includes an overview essay, a fairly linear historical essay on the fifth century B.C. that provides links into the primary material (e.g., Herodotus, Thucydides, and Plutarch), via small gray underscoring (on the screen) of linked passages. These are handmade links into other material.

To different extents, most of the production work was done at Harvard, where the people and the equipment are located. Much of the collaborative activity involved data collection and structuring, because the main challenge and the emphasis of Perseus is the gathering of primary material, that is, building a useful environment for studying

classical Greece, collecting data, and making it useful.

Systems-building is definitely not the main concern. Thus, much of the work has involved writing essays, collecting information, rewriting it, and tagging it. That can be done off site. The creative link for the overview essay as well as for both systems and data was collaborative, and was forged via E-mail and paper mail with professors at Pomona and Bowdoin.

******

++

CALALUCA * PLD's principal focus and contribution to scholarship * Various questions preparatory to beginning the project * Basis for project * Basic rule in converting PLD * Concerning the images in PLD * Running PLD under a variety of retrieval softwares * Encoding the database a hard-fought issue * Various features demonstrated * Importance of user documentation * Limitations of the CD-ROM version *

++

Eric CALALUCA, vice president, Chadwyck-Healey, Inc., demonstrated a software interpretation of the Patrologia Latina Database (PLD). PLD's principal focus from the beginning of the project about three-and-a-half years ago was on converting Migne's Latin series, and in the end, CALALUCA suggested, conversion of the text will be the major contribution to scholarship. CALALUCA stressed that, as possibly the only private publishing organization at the Workshop, Chadwyck-Healey had sought no

federal funds or national foundation support before embarking upon the project, but instead had relied upon a great deal of homework and marketing to accomplish the task of conversion.

Ever since the possibilities of computer-searching have emerged, scholars in the field of late ancient and early medieval studies (philosophers, theologians, classicists, and those studying the history of natural law and the history of the legal development of Western civilization) have been longing for a fully searchable version of Western literature, for example, all the texts of Augustine and Bernard of Clairvaux and Boethius, not to mention all the secondary and tertiary authors.

Various questions arose, CALALUCA said. Should one convert Migne? Should the database be encoded? Is it necessary to do that? How should it be delivered? What about CD-ROM? Since this is a transitional medium, why even bother to create software to run on a CD-ROM? Since everybody knows people will be networking information, why go to the trouble--which is far greater with CD-ROM than with the production of magnetic data? Finally, how does one make the data available? Can many of the hurdles to using electronic information that some publishers have imposed upon databases be eliminated?

The PLD project was based on the principle that computer-searching of texts is most effective when it is done with a large database. Because PLD represented a collection that serves so many disciplines across so many periods, it was irresistible.

The basic rule in converting PLD was to do no harm, to avoid the sins of intrusion in such a database: no introduction of newer editions, no on-the-spot changes, no eradicating of all possible falsehoods from an edition. Thus, PLD is not the final act in electronic publishing for this discipline, but simply the beginning. The conversion of PLD has evoked numerous unanticipated questions: How will information be used? What about networking? Can the rights of a database be protected? Should one protect the rights of a database? How can it be made available?

Those converting PLD also tried to avoid the sins of omission, that is, excluding portions of the collections or whole sections. What about the images? PLD is full of images, some are extremely pious nineteenth-century representations of the Fathers, while others contain highly interesting elements. The goal was to cover all the text of Migne (including notes, in Greek and in Hebrew, the latter of which, in particular, causes problems in creating a search structure), all the indices, and even the images, which are being scanned in separately searchable files.

Several North American institutions that have placed acquisition requests for the PLD database have requested it in magnetic form without software, which means they are already running it without software, without anything demonstrated at the Workshop.

What cannot practically be done is go back and reconvert and re-encode data, a time-consuming and extremely costly enterprise. CALALUCA sees PLD as a database that can, and should, be run under a variety of retrieval softwares. This will permit the widest possible searches. Consequently, the need to produce a CD-ROM of PLD, as well as to develop software that could handle some 1.3 gigabyte of heavily encoded text, developed out of conversations with collection development and reference librarians who wanted software both compassionate enough for the pedestrian but also capable of incorporating the most detailed lexicographical studies that a user desires to conduct. In the end, the encoding and conversion of the data will prove the most enduring testament to the value of the project.

The encoding of the database was also a hard-fought issue: Did the database need to be encoded? Were there normative structures for encoding humanist texts? Should it be SGML? What about the TEI--will it last, will it prove useful? CALALUCA expressed some minor doubts as to whether a data bank can be fully TEI-conformant. Every effort can be made, but in the end to be TEI-conformant means to accept the need to make some firm encoding decisions that can, indeed, be disputed. The TEI points the publisher in a proper direction but does not presume to make all the decisions for him or her. Essentially, the goal of encoding was to eliminate, as much as possible, the hindrances to information-networking, so that if an institution acquires a database, everybody associated with the institution can have access to it.

CALALUCA demonstrated a portion of Volume 160, because it had the most anomalies in it. The software was created by Electronic Book Technologies of Providence, RI, and is called Dynatext. The software works only with SGML-coded data.

Viewing a table of contents on the screen, the audience saw how Dynatext treats each element as a book and attempts to simplify movement through a volume. Familiarity with the Patrologia in print (i.e., the text, its source, and the editions) will make the machine-readable versions highly useful. (Software with a Windows application was sought for PLD, CALALUCA said, because this was the main trend for scholarly use.)

CALALUCA also demonstrated how a user can perform a variety of searches and quickly move to any part of a volume; the look-up screen provides some basic, simple word-searching.

CALALUCA argued that one of the major difficulties is not the software. Rather, in creating a product that will be used by scholars representing a broad spectrum of computer sophistication, user documentation proves to be the most important service one can provide.

CALALUCA next illustrated a truncated search under mysterium within ten words of virtus and how one would be able to find its contents throughout the entire database. He said that the exciting thing about PLD is that many of the applications in the retrieval software being written for it will exceed the capabilities of the software employed now for the CD-ROM

version. The CD-ROM faces genuine limitations, in terms of speed and comprehensiveness, in the creation of a retrieval software to run it.

CALALUCA said he hoped that individual scholars will download the data, if they wish, to their personal computers, and have ready access to important texts on a constant basis, which they will be able to use in their research and from which they might even be able to publish.

(CALALUCA explained that the blue numbers represented Migne's column numbers, which are the standard scholarly references. Pulling up a note, he stated that these texts were heavily edited and the image files would appear simply as a note as well, so that one could quickly access an image.)

******

++

FLEISCHHAUER/ERWAY * Several problems with which AM is still wrestling * Various search and retrieval capabilities * Illustration of automatic stemming and a truncated search * AM's attempt to find ways to connect cataloging to the texts * AM's gravitation towards SGML * Striking a balance between quantity and quality * How AM furnishes users recourse to images * Conducting a search in a full-text environment * Macintosh and IBM prototypes of AM * Multimedia aspects of AM *

++

A demonstration of American Memory by its coordinator, Carl FLEISCHHAUER, and Ricky ERWAY, associate coordinator, Library of Congress, concluded

the morning session. Beginning with a collection of broadsides from the Continental Congress and the Constitutional Convention, the only text collection in a presentable form at the time of the Workshop, FLEISCHHAUER highlighted several of the problems with which AM is still wrestling. (In its final form, the disk will contain two collections, not only the broadsides but also the full text with illustrations of a set of approximately 300 African-American pamphlets from the period 1870 to 1910.)

As FREEMAN had explained earlier, AM has attempted to use a small amount of interpretation to introduce collections. In the present case, the contractor, a company named Quick Source, in Silver Spring, MD., used software called Toolbook and put together a modestly interactive introduction to the collection. Like the two preceding speakers, FLEISCHHAUER argued that the real asset was the underlying collection.

FLEISCHHAUER proceeded to describe various search and retrieval capabilities while ERWAY worked the computer. In this particular package the "go to" pull-down allowed the user in effect to jump out of Toolbook, where the interactive program was located, and enter the third-party software used by AM for this text collection, which is called Personal Librarian. This was the Windows version of Personal Librarian, a software application put together by a company in Rockville, Md.

Since the broadsides came from the Revolutionary War period, a search was conducted using the words British or war, with the default operator reset as or. FLEISCHHAUER demonstrated both automatic stemming (which finds

other forms of the same root) and a truncated search. One of Personal Librarian's strongest features, the relevance ranking, was represented by a chart that indicated how often words being sought appeared in documents, with the one receiving the most "hits" obtaining the highest score. The "hit list" that is supplied takes the relevance ranking into account, making the first hit, in effect, the one the software has selected as the most relevant example.

While in the text of one of the broadside documents, FLEISCHHAUER remarked AM's attempt to find ways to connect cataloging to the texts, which it does in different ways in different manifestations. In the case shown, the cataloging was pasted on: AM took MARC records that were written as on-line records right into one of the Library's mainframe retrieval programs, pulled them out, and handed them off to the contractor, who massaged them somewhat to display them in the manner shown. One of AM's questions is, Does the cataloguing normally performed in the mainframe work in this context, or had AM ought to think through adjustments?

FLEISCHHAUER made the additional point that, as far as the text goes, AM has gravitated towards SGML (he pointed to the boldface in the upper part of the screen). Although extremely limited in its ability to translate or interpret SGML, Personal Librarian will furnish both bold and italics on screen; a fairly easy thing to do, but it is one of the ways in which SGML is useful.

Striking a balance between quantity and quality has been a major concern

of AM, with accuracy being one of the places where project staff have felt that less than 100-percent accuracy was not unacceptable. FLEISCHHAUER cited the example of the standard of the rekeying industry, namely 99.95 percent; as one service bureau informed him, to go from 99.95 to 100 percent would double the cost.

FLEISCHHAUER next demonstrated how AM furnishes users recourse to images, and at the same time recalled LESK's pointed question concerning the number of people who would look at those images and the number who would work only with the text. If the implication of LESK's question was sound, FLEISCHHAUER said, it raised the stakes for text accuracy and reduced the value of the strategy for images.

Contending that preservation is always a bugaboo, FLEISCHHAUER demonstrated several images derived from a scan of a preservation microfilm that AM had made. He awarded a grade of C at best, perhaps a C minus or a C plus, for how well it worked out. Indeed, the matter of learning if other people had better ideas about scanning in general, and, in particular, scanning from microfilm, was one of the factors that drove AM to attempt to think through the agenda for the Workshop. Skew, for example, was one of the issues that AM in its ignorance had not reckoned would prove so difficult.

Further, the handling of images of the sort shown, in a desktop computer environment, involved a considerable amount of zooming and scrolling. Ultimately, AM staff feel that perhaps the paper copy that is printed out

might be the most useful one, but they remain uncertain as to how much on-screen reading users will do.

Returning to the text, FLEISCHHAUER asked viewers to imagine a person who might be conducting a search in a full-text environment. With this scenario, he proceeded to illustrate other features of Personal Librarian that he considered helpful; for example, it provides the ability to notice words as one reads. Clicking the "include" button on the bottom of the search window pops the words that have been highlighted into the search. Thus, a user can refine the search as he or she reads, re-executing the search and continuing to find things in the quest for materials. This software not only contains relevance ranking, Boolean operators, and truncation, it also permits one to perform word algebra, so to say, where one puts two or three words in parentheses and links them with one Boolean operator and then a couple of words in another set of parentheses and asks for things within so many words of others.

Until they became acquainted recently with some of the work being done in classics, the AM staff had not realized that a large number of the projects that involve electronic texts were being done by people with a profound interest in language and linguistics. Their search strategies and thinking are oriented to those fields, as is shown in particular by the Perseus example. As amateur historians, the AM staff were thinking more of searching for concepts and ideas than for particular words. Obviously, FLEISCHHAUER conceded, searching for concepts and ideas and searching for words may be two rather closely related things.

While displaying several images, FLEISCHHAUER observed that the Macintosh prototype built by AM contains a greater diversity of formats. Echoing a previous speaker, he said that it was easier to stitch things together in the Macintosh, though it tended to be a little more anemic in search and retrieval. AM, therefore, increasingly has been investigating sophisticated retrieval engines in the IBM format.

FLEISCHHAUER demonstrated several additional examples of the prototype interfaces: One was AM's metaphor for the network future, in which a kind of reading-room graphic suggests how one would be able to go around to different materials. AM contains a large number of photographs in analog video form worked up from a videodisc, which enable users to make copies to print or incorporate in digital documents. A frame-grabber is built into the system, making it possible to bring an image into a window and digitize or print it out.

FLEISCHHAUER next demonstrated sound recording, which included texts. Recycled from a previous project, the collection included sixty 78-rpm phonograph records of political speeches that were made during and immediately after World War I. These constituted approximately three hours of audio, as AM has digitized it, which occupy 150 megabytes on a CD. Thus, they are considerably compressed. From the catalogue card, FLEISCHHAUER proceeded to a transcript of a speech with the audio available and with highlighted text following it as it played.
A photograph has been added and a transcription made.

Considerable value has been added beyond what the Library of Congress normally would do in cataloguing a sound recording, which raises several questions for AM concerning where to draw lines about how much value it can afford to add and at what point, perhaps, this becomes more than AM could reasonably do or reasonably wish to do. FLEISCHHAUER also demonstrated a motion picture. As FREEMAN had reported earlier, the motion picture materials have proved the most popular, not surprisingly. This says more about the medium, he thought, than about AM's presentation of it.

Because AM's goal was to bring together things that could be used by historians or by people who were curious about history, turn-of-the-century footage seemed to represent the most appropriate collections from the Library of Congress in motion pictures. These were the very first films made by Thomas Edison's company and some others at that time. The particular example illustrated was a Biograph film, brought in with a frame-grabber into a window. A single videodisc contains about fifty titles and pieces of film from that period, all of New York City. Taken together, AM believes, they provide an interesting documentary resource.

******

++

DISCUSSION * Using the frame-grabber in AM * Volume of material processed and to be processed * Purpose of AM within LC * Cataloguing and the

nature of AM's material * SGML coding and the question of quality versus quantity *

++

During the question-and-answer period that followed FLEISCHHAUER's presentation, several clarifications were made.

AM is bringing in motion pictures from a videodisc. The frame-grabber devices create a window on a computer screen, which permits users to digitize a single frame of the movie or one of the photographs. It produces a crude, rough-and-ready image that high school students can incorporate into papers, and that has worked very nicely in this way.

Commenting on FLEISCHHAUER's assertion that AM was looking more at searching ideas than words, MYLONAS argued that without words an idea does not exist. FLEISCHHAUER conceded that he ought to have articulated his point more clearly. MYLONAS stated that they were in fact both talking about the same thing. By searching for words and by forcing people to focus on the word, the Perseus Project felt that they would get them to the idea. The way one reviews results is tailored more to one kind of user than another.

Concerning the total volume of material that has been processed in this way, AM at this point has in retrievable form seven or eight collections, all of them photographic. In the Macintosh environment, for example, there probably are 35,000-40,000 photographs. The sound recordings

number sixty items. The broadsides number about 300 items. There are 500 political cartoons in the form of drawings. The motion pictures, as individual items, number sixty to seventy.

AM also has a manuscript collection, the life history portion of one of the federal project series, which will contain 2,900 individual documents, all first-person narratives. AM has in process about 350 African-American pamphlets, or about 12,000 printed pages for the period 1870-1910. Also in the works are some 4,000 panoramic photographs. AM has recycled a fair amount of the work done by LC's Prints and Photographs Division during the Library's optical disk pilot project in the 1980s. For example, a special division of LC has tooled up and thought through all the ramifications of electronic presentation of photographs. Indeed, they are wheeling them out in great barrel loads. The purpose of AM within the Library, it is hoped, is to catalyze several of the other special collection divisions which have no particular experience with, in some cases, mixed feelings about, an activity such as AM. Moreover, in many cases the divisions may be characterized as not only lacking experience in "electronifying" things but also in automated cataloguing. MARC cataloguing as practiced in the United States is heavily weighted toward the description of monograph and serial materials, but is much thinner when one enters the world of manuscripts and things that are held in the Library's music collection and other units. In response to a comment by LESK, that AM's material is very heavily photographic, and is so primarily because individual records have been made for each photograph, FLEISCHHAUER observed that an item-level catalog record exists, for example, for each photograph in the Detroit

Publishing collection of 25,000 pictures. In the case of the Federal Writers Project, for which nearly 3,000 documents exist, representing information from twenty-six different states, AM with the assistance of Karen STUART of the Manuscript Division will attempt to find some way not only to have a collection-level record but perhaps a MARC record for each state, which will then serve as an umbrella for the 100-200 documents that come under it. But that drama remains to be enacted. The AM staff is conservative and clings to cataloguing, though of course visitors tout artificial intelligence and neural networks in a manner that suggests that perhaps one need not have cataloguing or that much of it could be put aside.

The matter of SGML coding, FLEISCHHAUER conceded, returned the discussion to the earlier treated question of quality versus quantity in the Library of Congress. Of course, text conversion can be done with 100-percent accuracy, but it means that when one's holdings are as vast as LC's only a tiny amount will be exposed, whereas permitting lower levels of accuracy can lead to exposing or sharing larger amounts, but with the quality correspondingly impaired.

******

++

TWOHIG * A contrary experience concerning electronic options * Volume of material in the Washington papers and a suggestion of David Packard * Implications of Packard's suggestion * Transcribing the documents for the CD-ROM * Accuracy of transcriptions * The CD-ROM edition of the Founding

Fathers documents *

++

Finding encouragement in a comment of MICHELSON's from the morning session--that numerous people in the humanities were choosing electronic options to do their work--Dorothy TWOHIG, editor, The Papers of George Washington, opened her illustrated talk by noting that her experience with literary scholars and numerous people in editing was contrary to MICHELSON's. TWOHIG emphasized literary scholars' complete ignorance of the technological options available to them or their reluctance or, in some cases, their downright hostility toward these options.

After providing an overview of the five Founding Fathers projects (Jefferson at Princeton, Franklin at Yale, John Adams at the Massachusetts Historical Society, and Madison down the hall from her at the University of Virginia), TWOHIG observed that the Washington papers, like all of the projects, include both sides of the Washington correspondence and deal with some 135,000 documents to be published with extensive annotation in eighty to eighty-five volumes, a project that will not be completed until well into the next century. Thus, it was with considerable enthusiasm several years ago that the Washington Papers Project (WPP) greeted David Packard's suggestion that the papers of the Founding Fathers could be published easily and inexpensively, and to the great benefit of American scholarship, via CD-ROM.

In pragmatic terms, funding from the Packard Foundation would expedite

the transcription of thousands of documents waiting to be put on disk in the WPP offices. Further, since the costs of collecting, editing, and converting the Founding Fathers documents into letterpress editions were running into the millions of dollars, and the considerable staffs involved in all of these projects were devoting their careers to producing the work, the Packard Foundation's suggestion had a revolutionary aspect: Transcriptions of the entire corpus of the Founding Fathers papers would be available on CD-ROM to public and college libraries, even high schools, at a fraction of the cost--$100-$150 for the annual license fee--to produce a limited university press run of 1,000 of each volume of the published papers at $45-$150 per printed volume. Given the current budget crunch in educational systems and the corresponding constraints on librarians in smaller institutions who wish to add these volumes to their collections, producing the documents on CD-ROM would likely open a greatly expanded audience for the papers. TWOHIG stressed, however, that development of the Founding Fathers CD-ROM is still in its infancy. Serious software problems remain to be resolved before the material can be put into readable form.

Funding from the Packard Foundation resulted in a major push to transcribe the 75,000 or so documents of the Washington papers remaining to be transcribed onto computer disks. Slides illustrated several of the problems encountered, for example, the present inability of CD-ROM to indicate the cross-outs (deleted material) in eighteenth century documents. TWOHIG next described documents from various periods in the eighteenth century that have been transcribed in chronological order and delivered to the Packard offices in California, where they are converted

to the CD-ROM, a process that is expected to consume five years to complete (that is, reckoning from David Packard's suggestion made several years ago, until about July 1994). TWOHIG found an encouraging indication of the project's benefits in the ongoing use made by scholars of the search functions of the CD-ROM, particularly in reducing the time spent in manually turning the pages of the Washington papers.

TWOHIG next furnished details concerning the accuracy of transcriptions. For instance, the insertion of thousands of documents on the CD-ROM currently does not permit each document to be verified against the original manuscript several times as in the case of documents that appear in the published edition. However, the transcriptions receive a cursory check for obvious typos, the misspellings of proper names, and other errors from the WPP CD-ROM editor. Eventually, all documents that appear in the electronic version will be checked by project editors. Although this process has met with opposition from some of the editors on the grounds that imperfect work may leave their offices, the advantages in making this material available as a research tool outweigh fears about the misspelling of proper names and other relatively minor editorial matters.

Completion of all five Founding Fathers projects (i.e., retrievability and searchability of all of the documents by proper names, alternate spellings, or varieties of subjects) will provide one of the richest sources of this size for the history of the United States in the latter part of the eighteenth century. Further, publication on CD-ROM will allow editors to include even minutiae, such as laundry lists, not included in the printed volumes.

It seems possible that the extensive annotation provided in the printed volumes eventually will be added to the CD-ROM edition, pending negotiations with the publishers of the papers. At the moment, the Founding Fathers CD-ROM is accessible only on the IBYCUS, a computer developed out of the Thesaurus Linguae Graecae project and designed for the use of classical scholars. There are perhaps 400 IBYCUS computers in the country, most of which are in university classics departments. Ultimately, it is anticipated that the CD-ROM edition of the Founding Fathers documents will run on any IBM-compatible or Macintosh computer with a CD-ROM drive. Numerous changes in the software will also occur before the project is completed. (Editor's note: an IBYCUS was unavailable to demonstrate the CD-ROM.)

******

++

DISCUSSION * Several additional features of WPP clarified *

++

Discussion following TWOHIG's presentation served to clarify several additional features, including (1) that the project's primary intellectual product consists in the electronic transcription of the material; (2) that the text transmitted to the CD-ROM people is not marked up; (3) that cataloging and subject-indexing of the material remain to be worked out (though at this point material can be retrieved

by name); and (4) that because all the searching is done in the hardware, the IBYCUS is designed to read a CD-ROM which contains only sequential text files. Technically, it then becomes very easy to read the material off and put it on another device.

******

++

LEBRON * Overview of the history of the joint project between AAAS and OCLC * Several practices the on-line environment shares with traditional publishing on hard copy * Several technical and behavioral barriers to electronic publishing * How AAAS and OCLC arrived at the subject of clinical trials * Advantages of the electronic format and other features of OJCCT * An illustrated tour of the journal *

++

Maria LEBRON, managing editor, The Online Journal of Current Clinical Trials (OJCCT), presented an illustrated overview of the history of the joint project between the American Association for the Advancement of Science (AAAS) and the Online Computer Library Center, Inc. (OCLC). The joint venture between AAAS and OCLC owes its beginning to a reorganization launched by the new chief executive officer at OCLC about three years ago and combines the strengths of these two disparate organizations. In short, OJCCT represents the process of scholarly publishing on line.

LEBRON next discussed several practices the on-line environment shares with traditional publishing on hard copy--for example, peer review of manuscripts--that are highly important in the academic world. LEBRON noted in particular the implications of citation counts for tenure committees and grants committees. In the traditional hard-copy environment, citation counts are readily demonstrable, whereas the on-line environment represents an ethereal medium to most academics.

LEBRON remarked several technical and behavioral barriers to electronic publishing, for instance, the problems in transmission created by special characters or by complex graphics and halftones. In addition, she noted economic limitations such as the storage costs of maintaining back issues and market or audience education.

Manuscripts cannot be uploaded to OJCCT, LEBRON explained, because it is not a bulletin board or E-mail, forms of electronic transmission of information that have created an ambience clouding people's understanding of what the journal is attempting to do. OJCCT, which publishes peer-reviewed medical articles dealing with the subject of clinical trials, includes text, tabular material, and graphics, although at this time it can transmit only line illustrations.

Next, LEBRON described how AAAS and OCLC arrived at the subject of clinical trials: It is 1) a highly statistical discipline that 2) does not require halftones but can satisfy the needs of its audience with line illustrations and graphic material, and 3) there is a need for the speedy

dissemination of high-quality research results. Clinical trials are research activities that involve the administration of a test treatment to some experimental unit in order to test its usefulness before it is made available to the general population. LEBRON proceeded to give additional information on OJCCT concerning its editor-in-chief, editorial board, editorial content, and the types of articles it publishes (including peer-reviewed research reports and reviews), as well as features shared by other traditional hard-copy journals.

Among the advantages of the electronic format are faster dissemination of information, including raw data, and the absence of space constraints because pages do not exist. (This latter fact creates an interesting situation when it comes to citations.) Nor are there any issues. AAAS's capacity to download materials directly from the journal to a subscriber's printer, hard drive, or floppy disk helps ensure highly accurate transcription. Other features of OJCCT include on-screen alerts that allow linkage of subsequently published documents to the original documents; on-line searching by subject, author, title, etc.; indexing of every single word that appears in an article; viewing access to an article by component (abstract, full text, or graphs); numbered paragraphs to replace page counts; publication in Science every thirty days of indexing of all articles published in the journal; typeset-quality screens; and Hypertext links that enable subscribers to bring up Medline abstracts directly without leaving the journal.

After detailing the two primary ways to gain access to the journal, through the OCLC network and Compuserv if one desires graphics or through

the Internet if just an ASCII file is desired, LEBRON illustrated the
speedy editorial process and the coding of the document using SGML tags
after it has been accepted for publication. She also gave an illustrated
tour of the journal, its search-and-retrieval capabilities in particular,
but also including problems associated with scanning in illustrations,
and the importance of on-screen alerts to the medical profession re
retractions or corrections, or more frequently, editorials, letters to
the editors, or follow-up reports. She closed by inviting the audience
to join AAAS on 1 July, when OJCCT was scheduled to go on-line.

******

++

DISCUSSION * Additional features of OJCCT *

++

In the lengthy discussion that followed LEBRON's presentation, these
points emerged:

* The SGML text can be tailored as users wish.

* All these articles have a fairly simple document definition.

* Document-type definitions (DTDs) were developed and given to OJCCT
for coding.

* No articles will be removed from the journal. (Because there are no back issues, there are no lost issues either. Once a subscriber logs onto the journal he or she has access not only to the currently published materials, but retrospectively to everything that has been published in it. Thus the table of contents grows bigger. The date of publication serves to distinguish between currently published materials and older materials.)

* The pricing system for the journal resembles that for most medical journals: for 1992, $95 for a year, plus telecommunications charges (there are no connect time charges); for 1993, $110 for the entire year for single users, though the journal can be put on a local area network (LAN). However, only one person can access the journal at a time. Site licenses may come in the future.

* AAAS is working closely with colleagues at OCLC to display mathematical equations on screen.

* Without compromising any steps in the editorial process, the technology has reduced the time lag between when a manuscript is originally submitted and the time it is accepted; the review process does not differ greatly from the standard six-to-eight weeks employed by many of the hard-copy journals. The process still depends on people.

* As far as a preservation copy is concerned, articles will be maintained on the computer permanently and subscribers, as part of their subscription, will receive a microfiche-quality archival copy of everything published during that year; in addition, reprints can be purchased in much the same way as in a hard-copy environment. Hard copies are prepared but are not the primary medium for the dissemination of the information.

* Because OJCCT is not yet on line, it is difficult to know how many people would simply browse through the journal on the screen as opposed to downloading the whole thing and printing it out; a mix of both types of users likely will result.

******

++

PERSONIUS * Developments in technology over the past decade * The CLASS Project * Advantages for technology and for the CLASS Project * Developing a network application an underlying assumption of the project * Details of the scanning process * Print-on-demand copies of books * Future plans include development of a browsing tool *

++

Lynne PERSONIUS, assistant director, Cornell Information Technologies for Scholarly Information Services, Cornell University, first commented on

the tremendous impact that developments in technology over the past ten years--networking, in particular--have had on the way information is handled, and how, in her own case, these developments have counterbalanced Cornell's relative geographical isolation. Other significant technologies include scanners, which are much more sophisticated than they were ten years ago; mass storage and the dramatic savings that result from it in terms of both space and money relative to twenty or thirty years ago; new and improved printing technologies, which have greatly affected the distribution of information; and, of course, digital technologies, whose applicability to library preservation remains at issue.

Given that context, PERSONIUS described the College Library Access and Storage System (CLASS) Project, a library preservation project, primarily, and what has been accomplished. Directly funded by the Commission on Preservation and Access and by the Xerox Corporation, which has provided a significant amount of hardware, the CLASS Project has been working with a development team at Xerox to develop a software application tailored to library preservation requirements. Within Cornell, participants in the project have been working jointly with both library and information technologies. The focus of the project has been on reformatting and saving books that are in brittle condition. PERSONIUS showed Workshop participants a brittle book, and described how such books were the result of developments in papermaking around the beginning of the Industrial Revolution. The papermaking process was changed so that a significant amount of acid was introduced into the actual paper itself, which deteriorates as it sits on library shelves.

One of the advantages for technology and for the CLASS Project is that the information in brittle books is mostly out of copyright and thus offers an opportunity to work with material that requires library preservation, and to create and work on an infrastructure to save the material. Acknowledging the familiarity of those working in preservation with this information, PERSONIUS noted that several things are being done: the primary preservation technology used today is photocopying of brittle material. Saving the intellectual content of the material is the main goal. With microfilm copy, the intellectual content is preserved on the assumption that in the future the image can be reformatted in any other way that then exists.

An underlying assumption of the CLASS Project from the beginning was that it would develop a network application. Project staff scan books at a workstation located in the library, near the brittle material. An image-server filing system is located at a distance from that workstation, and a printer is located in another building. All of the materials digitized and stored on the image-filing system are cataloged in the on-line catalogue. In fact, a record for each of these electronic books is stored in the RLIN database so that a record exists of what is in the digital library throughout standard catalogue procedures. In the future, researchers working from their own workstations in their offices, or their networks, will have access--wherever they might be--through a request server being built into the new digital library. A second assumption is that the preferred means of finding the material will be by looking through a catalogue. PERSONIUS described the scanning process, which uses a prototype scanner being developed by Xerox and which scans a

very high resolution image at great speed. Another significant feature, because this is a preservation application, is the placing of the pages that fall apart one for one on the platen. Ordinarily, a scanner could be used with some sort of a document feeder, but because of this application that is not feasible. Further, because CLASS is a preservation application, after the paper replacement is made there, a very careful quality control check is performed. An original book is compared to the printed copy and verification is made, before proceeding, that all of the image, all of the information, has been captured. Then, a new library book is produced: The printed images are rebound by a commercial binder and a new book is returned to the shelf. Significantly, the books returned to the library shelves are beautiful and useful replacements on acid-free paper that should last a long time, in effect, the equivalent of preservation photocopies. Thus, the project has a library of digital books. In essence, CLASS is scanning and storing books as 600 dot-per-inch bit-mapped images, compressed using Group 4 CCITT (i.e., the French acronym for International Consultative Committee for Telegraph and Telephone) compression. They are stored as TIFF files on an optical filing system that is composed of a database used for searching and locating the books and an optical jukebox that stores 64 twelve-inch platters. A very-high-resolution printed copy of these books at 600 dots per inch is created, using a Xerox DocuTech printer to make the paper replacements on acid-free paper.

PERSONIUS maintained that the CLASS Project presents an opportunity to introduce people to books as digital images by using a paper medium. Books are returned to the shelves while people are also given the ability

to print on demand--to make their own copies of books. (PERSONIUS distributed copies of an engineering journal published by engineering students at Cornell around 1900 as an example of what a print-on-demand copy of material might be like. This very cheap copy would be available to people to use for their own research purposes and would bridge the gap between an electronic work and the paper that readers like to have.) PERSONIUS then attempted to illustrate a very early prototype of networked access to this digital library. Xerox Corporation has developed a prototype of a view station that can send images across the network to be viewed.

The particular library brought down for demonstration contained two mathematics books. CLASS is developing and will spend the next year developing an application that allows people at workstations to browse the books. Thus, CLASS is developing a browsing tool, on the assumption that users do not want to read an entire book from a workstation, but would prefer to be able to look through and decide if they would like to have a printed copy of it.

******

++

DISCUSSION * Re retrieval software * "Digital file copyright" * Scanning rate during production * Autosegmentation * Criteria employed in selecting books for scanning * Compression and decompression of images * OCR not precluded *

++

During the question-and-answer period that followed her presentation, PERSONIUS made these additional points:

* Re retrieval software, Cornell is developing a Unix-based server as well as clients for the server that support multiple platforms (Macintosh, IBM and Sun workstations), in the hope that people from any of those platforms will retrieve books; a further operating assumption is that standard interfaces will be used as much as possible, where standards can be put in place, because CLASS considers this retrieval software a library application and would like to be able to look at material not only at Cornell but at other institutions.

* The phrase "digital file copyright by Cornell University" was added at the advice of Cornell's legal staff with the caveat that it probably would not hold up in court. Cornell does not want people to copy its books and sell them but would like to keep them available for use in a library environment for library purposes.

* In production the scanner can scan about 300 pages per hour, capturing 600 dots per inch.

* The Xerox software has filters to scan halftone material and avoid

the moire patterns that occur when halftone material is scanned. Xerox has been working on hardware and software that would enable the scanner itself to recognize this situation and deal with it appropriately--a kind of autosegmentation that would enable the scanner to handle halftone material as well as text on a single page.

* The books subjected to the elaborate process described above were selected because CLASS is a preservation project, with the first 500 books selected coming from Cornell's mathematics collection, because they were still being heavily used and because, although they were in need of preservation, the mathematics library and the mathematics faculty were uncomfortable having them microfilmed. (They wanted a printed copy.) Thus, these books became a logical choice for this project. Other books were chosen by the project's selection committees for experiments with the technology, as well as to meet a demand or need.

* Images will be decompressed before they are sent over the line; at this time they are compressed and sent to the image filing system and then sent to the printer as compressed images; they are returned to the workstation as compressed 600-dpi images and the workstation decompresses and scales them for display--an inefficient way to access the material though it works quite well for printing and other purposes.

* CLASS is also decompressing on Macintosh and IBM, a slow process right now. Eventually, compression and decompression will take

place on an image conversion server. Trade-offs will be made, based on future performance testing, concerning where the file is compressed and what resolution image is sent.

* OCR has not been precluded; images are being stored that have been scanned at a high resolution, which presumably would suit them well to an OCR process. Because the material being scanned is about 100 years old and was printed with less-than-ideal technologies, very early and preliminary tests have not produced good results. But the project is capturing an image that is of sufficient resolution to be subjected to OCR in the future. Moreover, the system architecture and the system plan have a logical place to store an OCR image if it has been captured. But that is not being done now.

******

SESSION III. DISTRIBUTION, NETWORKS, AND NETWORKING: OPTIONS FOR DISSEMINATION

++

ZICH * Issues pertaining to CD-ROMs * Options for publishing in CD-ROM *

++

Robert ZICH, special assistant to the associate librarian for special projects, Library of Congress, and moderator of this session, first noted

the blessed but somewhat awkward circumstance of having four very distinguished people representing networks and networking or at least leaning in that direction, while lacking anyone to speak from the strongest possible background in CD-ROMs. ZICH expressed the hope that members of the audience would join the discussion. He stressed the subtitle of this particular session, "Options for Dissemination," and, concerning CD-ROMs, the importance of determining when it would be wise to consider dissemination in CD-ROM versus networks. A shopping list of issues pertaining to CD-ROMs included: the grounds for selecting commercial publishers, and in-house publication where possible versus nonprofit or government publication. A similar list for networks included: determining when one should consider dissemination through a network, identifying the mechanisms or entities that exist to place items on networks, identifying the pool of existing networks, determining how a producer would choose between networks, and identifying the elements of a business arrangement in a network.

Options for publishing in CD-ROM: an outside publisher versus self-publication. If an outside publisher is used, it can be nonprofit, such as the Government Printing Office (GPO) or the National Technical Information Service (NTIS), in the case of government. The pros and cons associated with employing an outside publisher are obvious. Among the pros, there is no trouble getting accepted. One pays the bill and, in effect, goes one's way. Among the cons, when one pays an outside publisher to perform the work, that publisher will perform the work it is obliged to do, but perhaps without the production expertise and skill in marketing and dissemination that some would seek. There is the body of

commercial publishers that do possess that kind of expertise in distribution and marketing but that obviously are selective. In self-publication, one exercises full control, but then one must handle matters such as distribution and marketing. Such are some of the options for publishing in the case of CD-ROM.

In the case of technical and design issues, which are also important, there are many matters which many at the Workshop already knew a good deal about: retrieval system requirements and costs, what to do about images, the various capabilities and platforms, the trade-offs between cost and performance, concerns about local-area networkability, interoperability, etc.

******

++

LYNCH * Creating networked information is different from using networks as an access or dissemination vehicle * Networked multimedia on a large scale does not yet work * Typical CD-ROM publication model a two-edged sword * Publishing information on a CD-ROM in the present world of immature standards * Contrast between CD-ROM and network pricing * Examples demonstrated earlier in the day as a set of insular information gems * Paramount need to link databases * Layering to become increasingly necessary * Project NEEDS and the issues of information reuse and active versus passive use * X-Windows as a way of differentiating between network access and networked information * Barriers to the distribution

of networked multimedia information * Need for good, real-time delivery protocols * The question of presentation integrity in client-server computing in the academic world * Recommendations for producing multimedia

++

Clifford LYNCH, director, Library Automation, University of California, opened his talk with the general observation that networked information constituted a difficult and elusive topic because it is something just starting to develop and not yet fully understood. LYNCH contended that creating genuinely networked information was different from using networks as an access or dissemination vehicle and was more sophisticated and more subtle. He invited the members of the audience to extrapolate, from what they heard about the preceding demonstration projects, to what sort of a world of electronics information--scholarly, archival, cultural, etc.--they wished to end up with ten or fifteen years from now. LYNCH suggested that to extrapolate directly from these projects would produce unpleasant results.

Putting the issue of CD-ROM in perspective before getting into generalities on networked information, LYNCH observed that those engaged in multimedia today who wish to ship a product, so to say, probably do not have much choice except to use CD-ROM: networked multimedia on a large scale basically does not yet work because the technology does not exist. For example, anybody who has tried moving images around over the Internet knows that this is an exciting touch-and-go process, a fascinating and fertile area for experimentation, research, and development, but not something that one can become deeply enthusiastic

about committing to production systems at this time.

This situation will change, LYNCH said. He differentiated CD-ROM from the practices that have been followed up to now in distributing data on CD-ROM. For LYNCH the problem with CD-ROM is not its portability or its slowness but the two-edged sword of having the retrieval application and the user interface inextricably bound up with the data, which is the typical CD-ROM publication model. It is not a case of publishing data but of distributing a typically stand-alone, typically closed system, all--software, user interface, and data--on a little disk. Hence, all the between-disk navigational issues as well as the impossibility in most cases of integrating data on one disk with that on another. Most CD-ROM retrieval software does not network very gracefully at present. However, in the present world of immature standards and lack of understanding of what network information is or what the ground rules are for creating or using it, publishing information on a CD-ROM does add value in a very real sense.

LYNCH drew a contrast between CD-ROM and network pricing and in doing so highlighted something bizarre in information pricing. A large institution such as the University of California has vendors who will offer to sell information on CD-ROM for a price per year in four digits, but for the same data (e.g., an abstracting and indexing database) on magnetic tape, regardless of how many people may use it concurrently, will quote a price in six digits.

What is packaged with the CD-ROM in one sense adds value--a complete access system, not just raw, unrefined information--although it is not generally perceived that way. This is because the access software, although it adds value, is viewed by some people, particularly in the university environment where there is a very heavy commitment to networking, as being developed in the wrong direction.

Given that context, LYNCH described the examples demonstrated as a set of insular information gems--Perseus, for example, offers nicely linked information, but would be very difficult to integrate with other databases, that is, to link together seamlessly with other source files from other sources. It resembles an island, and in this respect is similar to numerous stand-alone projects that are based on videodiscs, that is, on the single-workstation concept.

As scholarship evolves in a network environment, the paramount need will be to link databases. We must link personal databases to public databases, to group databases, in fairly seamless ways--which is extremely difficult in the environments under discussion with copies of databases proliferating all over the place.

The notion of layering also struck LYNCH as lurking in several of the projects demonstrated. Several databases in a sense constitute information archives without a significant amount of navigation built in. Educators, critics, and others will want a layered structure--one that defines or links paths through the layers to allow users to reach

specific points. In LYNCH's view, layering will become increasingly necessary, and not just within a single resource but across resources (e.g., tracing mythology and cultural themes across several classics databases as well as a database of Renaissance culture). This ability to organize resources, to build things out of multiple other things on the network or select pieces of it, represented for LYNCH one of the key aspects of network information.

Contending that information reuse constituted another significant issue, LYNCH commended to the audience's attention Project NEEDS (i.e., National Engineering Education Delivery System). This project's objective is to produce a database of engineering courseware as well as the components that can be used to develop new courseware. In a number of the existing applications, LYNCH said, the issue of reuse (how much one can take apart and reuse in other applications) was not being well considered. He also raised the issue of active versus passive use, one aspect of which is how much information will be manipulated locally by users. Most people, he argued, may do a little browsing and then will wish to print. LYNCH was uncertain how these resources would be used by the vast majority of users in the network environment.

LYNCH next said a few words about X-Windows as a way of differentiating between network access and networked information. A number of the applications demonstrated at the Workshop could be rewritten to use X across the network, so that one could run them from any X-capable device--a workstation, an X terminal--and transact with a database across the network. Although this opens up access a little, assuming one has enough

network to handle it, it does not provide an interface to develop a program that conveniently integrates information from multiple databases. X is a viewing technology that has limits. In a real sense, it is just a graphical version of remote log-in across the network. X-type applications represent only one step in the progression towards real access.

LYNCH next discussed barriers to the distribution of networked multimedia information. The heart of the problem is a lack of standards to provide the ability for computers to talk to each other, retrieve information, and shuffle it around fairly casually. At the moment, little progress is being made on standards for networked information; for example, present standards do not cover images, digital voice, and digital video. A useful tool kit of exchange formats for basic texts is only now being assembled. The synchronization of content streams (i.e., synchronizing a voice track to a video track, establishing temporal relations between different components in a multimedia object) constitutes another issue for networked multimedia that is just beginning to receive attention.

Underlying network protocols also need some work; good, real-time delivery protocols on the Internet do not yet exist. In LYNCH's view, highly important in this context is the notion of networked digital object IDs, the ability of one object on the network to point to another object (or component thereof) on the network. Serious bandwidth issues also exist. LYNCH was uncertain if billion-bit-per-second networks would prove sufficient if numerous people ran video in parallel.

LYNCH concluded by offering an issue for database creators to consider, as well as several comments about what might constitute good trial multimedia experiments. In a networked information world the database builder or service builder (publisher) does not exercise the same extensive control over the integrity of the presentation; strange programs "munge" with one's data before the user sees it. Serious thought must be given to what guarantees integrity of presentation. Part of that is related to where one draws the boundaries around a networked information service. This question of presentation integrity in client-server computing has not been stressed enough in the academic world, LYNCH argued, though commercial service providers deal with it regularly.

Concerning multimedia, LYNCH observed that good multimedia at the moment is hideously expensive to produce. He recommended producing multimedia with either very high sale value, or multimedia with a very long life span, or multimedia that will have a very broad usage base and whose costs therefore can be amortized among large numbers of users. In this connection, historical and humanistically oriented material may be a good place to start, because it tends to have a longer life span than much of the scientific material, as well as a wider user base. LYNCH noted, for example, that American Memory fits many of the criteria outlined. He remarked the extensive discussion about bringing the Internet or the National Research and Education Network (NREN) into the K-12 environment as a way of helping the American educational system.

LYNCH closed by noting that the kinds of applications demonstrated struck

him as excellent justifications of broad-scale networking for K-12, but that at this time no "killer" application exists to mobilize the K-12 community to obtain connectivity.

******

++

DISCUSSION * Dearth of genuinely interesting applications on the network a slow-changing situation * The issue of the integrity of presentation in a networked environment * Several reasons why CD-ROM software does not network *

++

During the discussion period that followed LYNCH's presentation, several additional points were made.

LYNCH reiterated even more strongly his contention that, historically, once one goes outside high-end science and the group of those who need access to supercomputers, there is a great dearth of genuinely interesting applications on the network. He saw this situation changing slowly, with some of the scientific databases and scholarly discussion groups and electronic journals coming on as well as with the availability of Wide Area Information Servers (WAIS) and some of the databases that are being mounted there. However, many of those things do not seem to have piqued great popular interest. For instance, most high school students of LYNCH's acquaintance would not qualify as devotees of serious

molecular biology.

Concerning the issue of the integrity of presentation, LYNCH believed that a couple of information providers have laid down the law at least on certain things. For example, his recollection was that the National Library of Medicine feels strongly that one needs to employ the identifier field if he or she is to mount a database commercially. The problem with a real networked environment is that one does not know who is reformatting and reprocessing one's data when one enters a client server mode. It becomes anybody's guess, for example, if the network uses a Z39.50 server, or what clients are doing with one's data. A data provider can say that his contract will only permit clients to have access to his data after he vets them and their presentation and makes certain it suits him. But LYNCH held out little expectation that the network marketplace would evolve in that way, because it required too much prior negotiation.

CD-ROM software does not network for a variety of reasons, LYNCH said. He speculated that CD-ROM publishers are not eager to have their products really hook into wide area networks, because they fear it will make their data suppliers nervous. Moreover, until relatively recently, one had to be rather adroit to run a full TCP/IP stack plus applications on a PC-size machine, whereas nowadays it is becoming easier as PCs grow bigger and faster. LYNCH also speculated that software providers had not heard from their customers until the last year or so, or had not heard from enough of their customers.

******

++

BESSER * Implications of disseminating images on the network; planning the distribution of multimedia documents poses two critical implementation problems * Layered approach represents the way to deal with users' capabilities * Problems in platform design; file size and its implications for networking * Transmission of megabyte size images impractical * Compression and decompression at the user's end * Promising trends for compression * A disadvantage of using X-Windows * A project at the Smithsonian that mounts images on several networks *

++

Howard BESSER, School of Library and Information Science, University of Pittsburgh, spoke primarily about multimedia, focusing on images and the broad implications of disseminating them on the network. He argued that planning the distribution of multimedia documents posed two critical implementation problems, which he framed in the form of two questions: 1) What platform will one use and what hardware and software will users have for viewing of the material? and 2) How can one deliver a sufficiently robust set of information in an accessible format in a reasonable amount of time? Depending on whether network or CD-ROM is the medium used, this question raises different issues of storage, compression, and transmission.

Concerning the design of platforms (e.g., sound, gray scale, simple color, etc.) and the various capabilities users may have, BESSER maintained that a layered approach was the way to deal with users' capabilities. A result would be that users with less powerful workstations would simply have less functionality. He urged members of the audience to advocate standards and accompanying software that handle layered functionality across a wide variety of platforms.

BESSER also addressed problems in platform design, namely, deciding how large a machine to design for situations when the largest number of users have the lowest level of the machine, and one desires higher functionality. BESSER then proceeded to the question of file size and its implications for networking. He discussed still images in the main. For example, a digital color image that fills the screen of a standard mega-pel workstation (Sun or Next) will require one megabyte of storage for an eight-bit image or three megabytes of storage for a true color or twenty-four-bit image. Lossless compression algorithms (that is, computational procedures in which no data is lost in the process of compressing [and decompressing] an image--the exact bit-representation is maintained) might bring storage down to a third of a megabyte per image, but not much further than that. The question of size makes it difficult to fit an appropriately sized set of these images on a single disk or to transmit them quickly enough on a network.

With these full screen mega-pel images that constitute a third of a megabyte, one gets 1,000-3,000 full-screen images on a one-gigabyte disk; a standard CD-ROM represents approximately 60 percent of that. Storing

images the size of a PC screen (just 8 bit color) increases storage capacity to 4,000-12,000 images per gigabyte; 60 percent of that gives one the size of a CD-ROM, which in turn creates a major problem. One cannot have full-screen, full-color images with lossless compression; one must compress them or use a lower resolution. For megabyte-size images, anything slower than a T-1 speed is impractical. For example, on a fifty-six-kilobaud line, it takes three minutes to transfer a one-megabyte file, if it is not compressed; and this speed assumes ideal circumstances (no other user contending for network bandwidth). Thus, questions of disk access, remote display, and current telephone connection speed make transmission of megabyte-size images impractical.

BESSER then discussed ways to deal with these large images, for example, compression and decompression at the user's end. In this connection, the issues of how much one is willing to lose in the compression process and what image quality one needs in the first place are unknown. But what is known is that compression entails some loss of data. BESSER urged that more studies be conducted on image quality in different situations, for example, what kind of images are needed for what kind of disciplines, and what kind of image quality is needed for a browsing tool, an intermediate viewing tool, and archiving.

BESSER remarked two promising trends for compression: from a technical perspective, algorithms that use what is called subjective redundancy employ principles from visual psycho-physics to identify and remove information from the image that the human eye cannot perceive; from an interchange and interoperability perspective, the JPEG (i.e., Joint

Photographic Experts Group, an ISO standard) compression algorithms also offer promise. These issues of compression and decompression, BESSER argued, resembled those raised earlier concerning the design of different platforms. Gauging the capabilities of potential users constitutes a primary goal. BESSER advocated layering or separating the images from the applications that retrieve and display them, to avoid tying them to particular software.

BESSER detailed several lessons learned from his work at Berkeley with Imagequery, especially the advantages and disadvantages of using X-Windows. In the latter category, for example, retrieval is tied directly to one's data, an intolerable situation in the long run on a networked system. Finally, BESSER described a project of Jim Wallace at the Smithsonian Institution, who is mounting images in a extremely rudimentary way on the Compuserv and Genie networks and is preparing to mount them on America On Line. Although the average user takes over thirty minutes to download these images (assuming a fairly fast modem), nevertheless, images have been downloaded 25,000 times.

BESSER concluded his talk with several comments on the business arrangement between the Smithsonian and Compuserv. He contended that not enough is known concerning the value of images.

******

++

DISCUSSION * Creating digitized photographic collections nearly impossible except with large organizations like museums * Need for study to determine quality of images users will tolerate *

++

During the brief exchange between LESK and BESSER that followed, several clarifications emerged.

LESK argued that the photographers were far ahead of BESSER: It is almost impossible to create such digitized photographic collections except with large organizations like museums, because all the photographic agencies have been going crazy about this and will not sign licensing agreements on any sort of reasonable terms. LESK had heard that National Geographic, for example, had tried to buy the right to use some image in some kind of educational production for $100 per image, but the photographers will not touch it. They want accounting and payment for each use, which cannot be accomplished within the system. BESSER responded that a consortium of photographers, headed by a former National Geographic photographer, had started assembling its own collection of electronic reproductions of images, with the money going back to the cooperative.

LESK contended that BESSER was unnecessarily pessimistic about multimedia images, because people are accustomed to low-quality images, particularly from video. BESSER urged the launching of a study to determine what users would tolerate, what they would feel comfortable with, and what

absolutely is the highest quality they would ever need. Conceding that he had adopted a dire tone in order to arouse people about the issue, BESSER closed on a sanguine note by saying that he would not be in this business if he did not think that things could be accomplished.

******

++

LARSEN * Issues of scalability and modularity * Geometric growth of the Internet and the role played by layering * Basic functions sustaining this growth * A library's roles and functions in a network environment * Effects of implementation of the Z39.50 protocol for information retrieval on the library system * The trade-off between volumes of data and its potential usage * A snapshot of current trends *

++

Ronald LARSEN, associate director for information technology, University of Maryland at College Park, first addressed the issues of scalability and modularity. He noted the difficulty of anticipating the effects of orders-of-magnitude growth, reflecting on the twenty years of experience with the Arpanet and Internet. Recalling the day's demonstrations of CD-ROM and optical disk material, he went on to ask if the field has yet learned how to scale new systems to enable delivery and dissemination across large-scale networks.

LARSEN focused on the geometric growth of the Internet from its inception

circa 1969 to the present, and the adjustments required to respond to that rapid growth. To illustrate the issue of scalability, LARSEN considered computer networks as including three generic components: computers, network communication nodes, and communication media. Each component scales (e.g., computers range from PCs to supercomputers; network nodes scale from interface cards in a PC through sophisticated routers and gateways; and communication media range from 2,400-baud dial-up facilities through 4.5-Mbps backbone links, and eventually to multigigabit-per-second communication lines), and architecturally, the components are organized to scale hierarchically from local area networks to international-scale networks. Such growth is made possible by building layers of communication protocols, as BESSER pointed out. By layering both physically and logically, a sense of scalability is maintained from local area networks in offices, across campuses, through bridges, routers, campus backbones, fiber-optic links, etc., up into regional networks and ultimately into national and international networks.

LARSEN then illustrated the geometric growth over a two-year period--through September 1991--of the number of networks that comprise the Internet. This growth has been sustained largely by the availability of three basic functions: electronic mail, file transfer (ftp), and remote log-on (telnet). LARSEN also reviewed the growth in the kind of traffic that occurs on the network. Network traffic reflects the joint contributions of a larger population of users and increasing use per user. Today one sees serious applications involving moving images across the network--a rarity ten years ago. LARSEN recalled and concurred with BESSER's main point

that the interesting problems occur at the application level.

LARSEN then illustrated a model of a library's roles and functions in a network environment. He noted, in particular, the placement of on-line catalogues onto the network and patrons obtaining access to the library increasingly through local networks, campus networks, and the Internet. LARSEN supported LYNCH's earlier suggestion that we need to address fundamental questions of networked information in order to build environments that scale in the information sense as well as in the physical sense.

LARSEN supported the role of the library system as the access point into the nation's electronic collections. Implementation of the Z39.50 protocol for information retrieval would make such access practical and feasible. For example, this would enable patrons in Maryland to search California libraries, or other libraries around the world that are conformant with Z39.50 in a manner that is familiar to University of Maryland patrons. This client-server model also supports moving beyond secondary content into primary content. (The notion of how one links from secondary content to primary content, LARSEN said, represents a fundamental problem that requires rigorous thought.) After noting numerous network experiments in accessing full-text materials, including projects supporting the ordering of materials across the network, LARSEN revisited the issue of transmitting high-density, high-resolution color images across the network and the large amounts of bandwidth they require. He went on to address the bandwidth and synchronization problems inherent in sending full-motion video across the network.

LARSEN illustrated the trade-off between volumes of data in bytes or orders of magnitude and the potential usage of that data. He discussed transmission rates (particularly, the time it takes to move various forms of information), and what one could do with a network supporting multigigabit-per-second transmission. At the moment, the network environment includes a composite of data-transmission requirements, volumes and forms, going from steady to bursty (high-volume) and from very slow to very fast. This aggregate must be considered in the design, construction, and operation of multigigabyte networks.

LARSEN's objective is to use the networks and library systems now being constructed to increase access to resources wherever they exist, and thus, to evolve toward an on-line electronic virtual library.

LARSEN concluded by offering a snapshot of current trends: continuing geometric growth in network capacity and number of users; slower development of applications; and glacial development and adoption of standards. The challenge is to design and develop each new application system with network access and scalability in mind.

******

++

BROWNRIGG * Access to the Internet cannot be taken for granted * Packet

radio and the development of MELVYL in 1980-81 in the Division of Library Automation at the University of California * Design criteria for packet radio * A demonstration project in San Diego and future plans * Spread spectrum * Frequencies at which the radios will run and plans to reimplement the WAIS server software in the public domain * Need for an infrastructure of radios that do not move around *

++

Edwin BROWNRIGG, executive director, Memex Research Institute, first polled the audience in order to seek out regular users of the Internet as well as those planning to use it some time in the future. With nearly everybody in the room falling into one category or the other, BROWNRIGG made a point re access, namely that numerous individuals, especially those who use the Internet every day, take for granted their access to it, the speeds with which they are connected, and how well it all works. However, as BROWNRIGG discovered between 1987 and 1989 in Australia, if one wants access to the Internet but cannot afford it or has some physical boundary that prevents her or him from gaining access, it can be extremely frustrating. He suggested that because of economics and physical barriers we were beginning to create a world of haves and have-nots in the process of scholarly communication, even in the United States.

BROWNRIGG detailed the development of MELVYL in academic year 1980-81 in the Division of Library Automation at the University of California, in order to underscore the issue of access to the system, which at the outset was extremely limited. In short, the project needed to build a network, which at that time entailed use of satellite technology, that is,

putting earth stations on campus and also acquiring some terrestrial links from the State of California's microwave system. The installation of satellite links, however, did not solve the problem (which actually formed part of a larger problem involving politics and financial resources). For while the project team could get a signal onto a campus, it had no means of distributing the signal throughout the campus. The solution involved adopting a recent development in wireless communication called packet radio, which combined the basic notion of packet-switching with radio. The project used this technology to get the signal from a point on campus where it came down, an earth station for example, into the libraries, because it found that wiring the libraries, especially the older marble buildings, would cost $2,000-$5,000 per terminal.

BROWNRIGG noted that, ten years ago, the project had neither the public policy nor the technology that would have allowed it to use packet radio in any meaningful way. Since then much had changed. He proceeded to detail research and development of the technology, how it is being deployed in California, and what direction he thought it would take. The design criteria are to produce a high-speed, one-time, low-cost, high-quality, secure, license-free device (packet radio) that one can plug in and play today, forget about it, and have access to the Internet. By high speed, BROWNRIGG meant 1 megabyte and 1.5 megabytes. Those units have been built, he continued, and are in the process of being type-certified by an independent underwriting laboratory so that they can be type-licensed by the Federal Communications Commission. As is the case with citizens band, one will be able to purchase a unit and not have to worry about applying for a license.

The basic idea, BROWNRIGG elaborated, is to take high-speed radio data transmission and create a backbone network that at certain strategic points in the network will "gateway" into a medium-speed packet radio (i.e., one that runs at 38.4 kilobytes), so that perhaps by 1994-1995 people, like those in the audience for the price of a VCR could purchase a medium-speed radio for the office or home, have full network connectivity to the Internet, and partake of all its services, with no need for an FCC license and no regular bill from the local common carrier. BROWNRIGG presented several details of a demonstration project currently taking place in San Diego and described plans, pending funding, to install a full-bore network in the San Francisco area. This network will have 600 nodes running at backbone speeds, and 100 of these nodes will be libraries, which in turn will be the gateway ports to the 38.4 kilobyte radios that will give coverage for the neighborhoods surrounding the libraries.

BROWNRIGG next explained Part 15.247, a new rule within Title 47 of the Code of Federal Regulations enacted by the FCC in 1985. This rule challenged the industry, which has only now risen to the occasion, to build a radio that would run at no more than one watt of output power and use a fairly exotic method of modulating the radio wave called spread spectrum. Spread spectrum in fact permits the building of networks so that numerous data communications can occur simultaneously, without interfering with each other, within the same wide radio channel.

BROWNRIGG explained that the frequencies at which the radios would run

are very short wave signals. They are well above standard microwave and radar. With a radio wave that small, one watt becomes a tremendous punch per bit and thus makes transmission at reasonable speed possible. In order to minimize the potential for congestion, the project is undertaking to reimplement software which has been available in the networking business and is taken for granted now, for example, TCP/IP, routing algorithms, bridges, and gateways. In addition, the project plans to take the WAIS server software in the public domain and reimplement it so that one can have a WAIS server on a Mac instead of a Unix machine. The Memex Research Institute believes that libraries, in particular, will want to use the WAIS servers with packet radio. This project, which has a team of about twelve people, will run through 1993 and will include the 100 libraries already mentioned as well as other professionals such as those in the medical profession, engineering, and law. Thus, the need is to create an infrastructure of radios that do not move around, which, BROWNRIGG hopes, will solve a problem not only for libraries but for individuals who, by and large today, do not have access to the Internet from their homes and offices.

******

++

DISCUSSION * Project operating frequencies *

++

During a brief discussion period, which also concluded the day's

proceedings, BROWNRIGG stated that the project was operating in four frequencies. The slow speed is operating at 435 megahertz, and it would later go up to 920 megahertz. With the high-speed frequency, the one-megabyte radios will run at 2.4 gigabits, and 1.5 will run at 5.7. At 5.7, rain can be a factor, but it would have to be tropical rain, unlike what falls in most parts of the United States.

******

SESSION IV. IMAGE CAPTURE, TEXT CAPTURE, OVERVIEW OF TEXT AND IMAGE STORAGE FORMATS

William HOOTON, vice president of operations, I-NET, moderated this session.

++

KENNEY * Factors influencing development of CXP * Advantages of using digital technology versus photocopy and microfilm * A primary goal of CXP; publishing challenges * Characteristics of copies printed * Quality of samples achieved in image capture * Several factors to be considered in choosing scanning * Emphasis of CXP on timely and cost-effective production of black-and-white printed facsimiles * Results of producing microfilm from digital files * Advantages of creating microfilm * Details concerning production * Costs * Role of digital technology in library preservation *

++

Anne KENNEY, associate director, Department of Preservation and Conservation, Cornell University, opened her talk by observing that the Cornell Xerox Project (CXP) has been guided by the assumption that the ability to produce printed facsimiles or to replace paper with paper would be important, at least for the present generation of users and equipment. She described three factors that influenced development of the project: 1) Because the project has emphasized the preservation of deteriorating brittle books, the quality of what was produced had to be sufficiently high to return a paper replacement to the shelf. CXP was only interested in using: 2) a system that was cost-effective, which meant that it had to be cost-competitive with the processes currently available, principally photocopy and microfilm, and 3) new or currently available product hardware and software.

KENNEY described the advantages that using digital technology offers over both photocopy and microfilm: 1) The potential exists to create a higher quality reproduction of a deteriorating original than conventional light-lens technology. 2) Because a digital image is an encoded representation, it can be reproduced again and again with no resulting loss of quality, as opposed to the situation with light-lens processes, in which there is discernible difference between a second and a subsequent generation of an image. 3) A digital image can be manipulated in a number of ways to improve image capture; for example, Xerox has developed a windowing application that enables one to capture a page containing both text and illustrations in a manner that optimizes the reproduction of both. (With light-lens technology, one must choose which

to optimize, text or the illustration; in preservation microfilming, the current practice is to shoot an illustrated page twice, once to highlight the text and the second time to provide the best capture for the illustration.) 4) A digital image can also be edited, density levels adjusted to remove underlining and stains, and to increase legibility for faint documents. 5) On-screen inspection can take place at the time of initial setup and adjustments made prior to scanning, factors that substantially reduce the number of retakes required in quality control.

A primary goal of CXP has been to evaluate the paper output printed on the Xerox DocuTech, a high-speed printer that produces 600-dpi pages from scanned images at a rate of 135 pages a minute. KENNEY recounted several publishing challenges to represent faithful and legible reproductions of the originals that the 600-dpi copy for the most part successfully captured. For example, many of the deteriorating volumes in the project were heavily illustrated with fine line drawings or halftones or came in languages such as Japanese, in which the buildup of characters comprised of varying strokes is difficult to reproduce at lower resolutions; a surprising number of them came with annotations and mathematical formulas, which it was critical to be able to duplicate exactly.

KENNEY noted that 1) the copies are being printed on paper that meets the ANSI standards for performance, 2) the DocuTech printer meets the machine and toner requirements for proper adhesion of print to page, as described by the National Archives, and thus 3) paper product is considered to be the archival equivalent of preservation photocopy.

KENNEY then discussed several samples of the quality achieved in the project that had been distributed in a handout, for example, a copy of a print-on-demand version of the 1911 Reed lecture on the steam turbine, which contains halftones, line drawings, and illustrations embedded in text; the first four loose pages in the volume compared the capture capabilities of scanning to photocopy for a standard test target, the IEEE standard 167A 1987 test chart. In all instances scanning proved superior to photocopy, though only slightly more so in one.

Conceding the simplistic nature of her review of the quality of scanning to photocopy, KENNEY described it as one representation of the kinds of settings that could be used with scanning capabilities on the equipment CXP uses. KENNEY also pointed out that CXP investigated the quality achieved with binary scanning only, and noted the great promise in gray scale and color scanning, whose advantages and disadvantages need to be examined. She argued further that scanning resolutions and file formats can represent a complex trade-off between the time it takes to capture material, file size, fidelity to the original, and on-screen display; and printing and equipment availability. All these factors must be taken into consideration.

CXP placed primary emphasis on the production in a timely and cost-effective manner of printed facsimiles that consisted largely of black-and-white text. With binary scanning, large files may be compressed efficiently and in a lossless manner (i.e., no data is lost in the process of compressing [and decompressing] an image--the exact

bit-representation is maintained) using Group 4 CCITT (i.e., the French acronym for International Consultative Committee for Telegraph and Telephone) compression. CXP was getting compression ratios of about forty to one. Gray-scale compression, which primarily uses JPEG, is much less economical and can represent a lossy compression (i.e., not lossless), so that as one compresses and decompresses, the illustration is subtly changed. While binary files produce a high-quality printed version, it appears 1) that other combinations of spatial resolution with gray and/or color hold great promise as well, and 2) that gray scale can represent a tremendous advantage for on-screen viewing. The quality associated with binary and gray scale also depends on the equipment used. For instance, binary scanning produces a much better copy on a binary printer.

Among CXP's findings concerning the production of microfilm from digital files, KENNEY reported that the digital files for the same Reed lecture were used to produce sample film using an electron beam recorder. The resulting film was faithful to the image capture of the digital files, and while CXP felt that the text and image pages represented in the Reed lecture were superior to that of the light-lens film, the resolution readings for the 600 dpi were not as high as standard microfilming. KENNEY argued that the standards defined for light-lens technology are not totally transferable to a digital environment. Moreover, they are based on definition of quality for a preservation copy. Although making this case will prove to be a long, uphill struggle, CXP plans to continue to investigate the issue over the course of the next year.

KENNEY concluded this portion of her talk with a discussion of the advantages of creating film: it can serve as a primary backup and as a preservation master to the digital file; it could then become the print or production master and service copies could be paper, film, optical disks, magnetic media, or on-screen display.

Finally, KENNEY presented details re production:

* Development and testing of a moderately-high resolution production scanning workstation represented a third goal of CXP; to date, 1,000 volumes have been scanned, or about 300,000 images.

* The resulting digital files are stored and used to produce hard-copy replacements for the originals and additional prints on demand; although the initial costs are high, scanning technology offers an affordable means for reformatting brittle material.

* A technician in production mode can scan 300 pages per hour when performing single-sheet scanning, which is a necessity when working with truly brittle paper; this figure is expected to increase significantly with subsequent iterations of the software from Xerox; a three-month time-and-cost study of scanning found that the average 300-page book would take about an hour and forty minutes to scan (this figure included the time for setup, which involves keying in primary bibliographic data, going into quality control mode to

define page size, establishing front-to-back registration, and scanning sample pages to identify a default range of settings for the entire book--functions not dissimilar to those performed by filmers or those preparing a book for photocopy).

* The final step in the scanning process involved rescans, which happily were few and far between, representing well under 1 percent of the total pages scanned.

In addition to technician time, CXP costed out equipment, amortized over four years, the cost of storing and refreshing the digital files every four years, and the cost of printing and binding, book-cloth binding, a paper reproduction. The total amounted to a little under $65 per single 300-page volume, with 30 percent overhead included--a figure competitive with the prices currently charged by photocopy vendors.

Of course, with scanning, in addition to the paper facsimile, one is left with a digital file from which subsequent copies of the book can be produced for a fraction of the cost of photocopy, with readers afforded choices in the form of these copies.

KENNEY concluded that digital technology offers an electronic means for a library preservation effort to pay for itself. If a brittle-book program included the means of disseminating reprints of books that are in demand by libraries and researchers alike, the initial investment in capture could be recovered and used to preserve additional but less popular

books. She disclosed that an economic model for a self-sustaining program could be developed for CXP's report to the Commission on Preservation and Access (CPA).

KENNEY stressed that the focus of CXP has been on obtaining high quality in a production environment. The use of digital technology is viewed as an affordable alternative to other reformatting options.

******

++

ANDRE * Overview and history of NATDP * Various agricultural CD-ROM products created inhouse and by service bureaus * Pilot project on Internet transmission * Additional products in progress *

++

Pamela ANDRE, associate director for automation, National Agricultural Text Digitizing Program (NATDP), National Agricultural Library (NAL), presented an overview of NATDP, which has been underway at NAL the last four years, before Judith ZIDAR discussed the technical details. ANDRE defined agricultural information as a broad range of material going from basic and applied research in the hard sciences to the one-page pamphlets that are distributed by the cooperative state extension services on such things as how to grow blueberries.

NATDP began in late 1986 with a meeting of representatives from the land-grant library community to deal with the issue of electronic information. NAL and forty-five of these libraries banded together to establish this project--to evaluate the technology for converting what were then source documents in paper form into electronic form, to provide access to that digital information, and then to distribute it. Distributing that material to the community--the university community as well as the extension service community, potentially down to the county level--constituted the group's chief concern.

Since January 1988 (when the microcomputer-based scanning system was installed at NAL), NATDP has done a variety of things, concerning which ZIDAR would provide further details. For example, the first technology considered in the project's discussion phase was digital videodisc, which indicates how long ago it was conceived.

Over the four years of this project, four separate CD-ROM products on four different agricultural topics were created, two at a scanning-and-OCR station installed at NAL, and two by service bureaus. Thus, NATDP has gained comparative information in terms of those relative costs. Each of these products contained the full ASCII text as well as page images of the material, or between 4,000 and 6,000 pages of material on these disks. Topics included aquaculture, food, agriculture and science (i.e., international agriculture and research), acid rain, and Agent Orange, which was the final product distributed (approximately eighteen months before the Workshop).

The third phase of NATDP focused on delivery mechanisms other than CD-ROM. At the suggestion of Clifford LYNCH, who was a technical consultant to the project at this point, NATDP became involved with the Internet and initiated a project with the help of North Carolina State University, in which fourteen of the land-grant university libraries are transmitting digital images over the Internet in response to interlibrary loan requests--a topic for another meeting. At this point, the pilot project had been completed for about a year and the final report would be available shortly after the Workshop. In the meantime, the project's success had led to its extension. (ANDRE noted that one of the first things done under the program title was to select a retrieval package to use with subsequent products; Windows Personal Librarian was the package of choice after a lengthy evaluation.)

Three additional products had been planned and were in progress:

1) An arrangement with the American Society of Agronomy--a professional society that has published the Agronomy Journal since about 1908--to scan and create bit-mapped images of its journal. ASA granted permission first to put and then to distribute this material in electronic form, to hold it at NAL, and to use these electronic images as a mechanism to deliver documents or print out material for patrons, among other uses. Effectively, NAL has the right to use this material in support of its program. (Significantly, this arrangement offers a potential cooperative model for working with other professional societies in agriculture to try to do the same thing--put the journals of particular interest

to agriculture research into electronic form.)

2) An extension of the earlier product on aquaculture.

3) The George Washington Carver Papers--a joint project with
Tuskegee University to scan and convert from microfilm some 3,500
images of Carver's papers, letters, and drawings.

It was anticipated that all of these products would appear no more than
six months after the Workshop.

******

++

ZIDAR * (A separate arena for scanning) * Steps in creating a database *
Image capture, with and without performing OCR * Keying in tracking data
* Scanning, with electronic and manual tracking * Adjustments during
scanning process * Scanning resolutions * Compression * De-skewing and
filtering * Image capture from microform: the papers and letters of
George Washington Carver * Equipment used for a scanning system *

++

Judith ZIDAR, coordinator, National Agricultural Text Digitizing Program
(NATDP), National Agricultural Library (NAL), illustrated the technical
details of NATDP, including her primary responsibility, scanning and

creating databases on a topic and putting them on CD-ROM.

(ZIDAR remarked a separate arena from the CD-ROM projects, although the processing of the material is nearly identical, in which NATDP is also scanning material and loading it on a Next microcomputer, which in turn is linked to NAL's integrated library system. Thus, searches in NAL's bibliographic database will enable people to pull up actual page images and text for any documents that have been entered.)

In accordance with the session's topic, ZIDAR focused her illustrated talk on image capture, offering a primer on the three main steps in the process: 1) assemble the printed publications; 2) design the database (database design occurs in the process of preparing the material for scanning; this step entails reviewing and organizing the material, defining the contents--what will constitute a record, what kinds of fields will be captured in terms of author, title, etc.); 3) perform a certain amount of markup on the paper publications. NAL performs this task record by record, preparing work sheets or some other sort of tracking material and designing descriptors and other enhancements to be added to the data that will not be captured from the printed publication. Part of this process also involves determining NATDP's file and directory structure: NATDP attempts to avoid putting more than approximately 100 images in a directory, because placing more than that on a CD-ROM would reduce the access speed.

This up-front process takes approximately two weeks for a

6,000-7,000-page database. The next step is to capture the page images. How long this process takes is determined by the decision whether or not to perform OCR. Not performing OCR speeds the process, whereas text capture requires greater care because of the quality of the image: it has to be straighter and allowance must be made for text on a page, not just for the capture of photographs.

NATDP keys in tracking data, that is, a standard bibliographic record including the title of the book and the title of the chapter, which will later either become the access information or will be attached to the front of a full-text record so that it is searchable.

Images are scanned from a bound or unbound publication, chiefly from bound publications in the case of NATDP, however, because often they are the only copies and the publications are returned to the shelves. NATDP usually scans one record at a time, because its database tracking system tracks the document in that way and does not require further logical separating of the images. After performing optical character recognition, NATDP moves the images off the hard disk and maintains a volume sheet. Though the system tracks electronically, all the processing steps are also tracked manually with a log sheet.

ZIDAR next illustrated the kinds of adjustments that one can make when scanning from paper and microfilm, for example, redoing images that need special handling, setting for dithering or gray scale, and adjusting for brightness or for the whole book at one time.

NATDP is scanning at 300 dots per inch, a standard scanning resolution. Though adequate for capturing text that is all of a standard size, 300 dpi is unsuitable for any kind of photographic material or for very small text. Many scanners allow for different image formats, TIFF, of course, being a de facto standard. But if one intends to exchange images with other people, the ability to scan other image formats, even if they are less common, becomes highly desirable.

CCITT Group 4 is the standard compression for normal black-and-white images, JPEG for gray scale or color. ZIDAR recommended 1) using the standard compressions, particularly if one attempts to make material available and to allow users to download images and reuse them from CD-ROMs; and 2) maintaining the ability to output an uncompressed image, because in image exchange uncompressed images are more likely to be able to cross platforms.

ZIDAR emphasized the importance of de-skewing and filtering as requirements on NATDP's upgraded system. For instance, scanning bound books, particularly books published by the federal government whose pages are skewed, and trying to scan them straight if OCR is to be performed, is extremely time-consuming. The same holds for filtering of poor-quality or older materials.

ZIDAR described image capture from microform, using as an example three reels from a sixty-seven-reel set of the papers and letters of George

Washington Carver that had been produced by Tuskegee University. These resulted in approximately 3,500 images, which NATDP had had scanned by its service contractor, Science Applications International Corporation (SAIC). NATDP also created bibliographic records for access. (NATDP did not have such specialized equipment as a microfilm scanner.

Unfortunately, the process of scanning from microfilm was not an unqualified success, ZIDAR reported: because microfilm frame sizes vary, occasionally some frames were missed, which without spending much time and money could not be recaptured.

OCR could not be performed from the scanned images of the frames. The bleeding in the text simply output text, when OCR was run, that could not even be edited. NATDP tested for negative versus positive images, landscape versus portrait orientation, and single- versus dual-page microfilm, none of which seemed to affect the quality of the image; but also on none of them could OCR be performed.

In selecting the microfilm they would use, therefore, NATDP had other factors in mind. ZIDAR noted two factors that influenced the quality of the images: 1) the inherent quality of the original and 2) the amount of size reduction on the pages.

The Carver papers were selected because they are informative and visually interesting, treat a single subject, and are valuable in their own right.
The images were scanned and divided into logical records by SAIC, then

delivered, and loaded onto NATDP's system, where bibliographic information taken directly from the images was added. Scanning was completed in summer 1991 and by the end of summer 1992 the disk was scheduled to be published.

Problems encountered during processing included the following: Because the microfilm scanning had to be done in a batch, adjustment for individual page variations was not possible. The frame size varied on account of the nature of the material, and therefore some of the frames were missed while others were just partial frames. The only way to go back and capture this material was to print out the page with the microfilm reader from the missing frame and then scan it in from the page, which was extremely time-consuming. The quality of the images scanned from the printout of the microfilm compared unfavorably with that of the original images captured directly from the microfilm. The inability to perform OCR also was a major disappointment. At the time, computer output microfilm was unavailable to test.

The equipment used for a scanning system was the last topic addressed by ZIDAR. The type of equipment that one would purchase for a scanning system included: a microcomputer, at least a 386, but preferably a 486; a large hard disk, 380 megabyte at minimum; a multi-tasking operating system that allows one to run some things in batch in the background while scanning or doing text editing, for example, Unix or OS/2 and, theoretically, Windows; a high-speed scanner and scanning software that allows one to make the various adjustments mentioned earlier; a high-resolution monitor (150 dpi); OCR software and hardware to perform

text recognition; an optical disk subsystem on which to archive all the images as the processing is done; file management and tracking software.

ZIDAR opined that the software one purchases was more important than the hardware and might also cost more than the hardware, but it was likely to prove critical to the success or failure of one's system. In addition to a stand-alone scanning workstation for image capture, then, text capture requires one or two editing stations networked to this scanning station to perform editing. Editing the text takes two or three times as long as capturing the images.

Finally, ZIDAR stressed the importance of buying an open system that allows for more than one vendor, complies with standards, and can be upgraded.

******

++

WATERS *Yale University Library's master plan to convert microfilm to digital imagery (POB) * The place of electronic tools in the library of the future * The uses of images and an image library * Primary input from preservation microfilm * Features distinguishing POB from CXP and key hypotheses guiding POB * Use of vendor selection process to facilitate organizational work * Criteria for selecting vendor * Finalists and results of process for Yale * Key factor distinguishing vendors * Components, design principles, and some estimated costs of POB * Role of preservation materials in developing imaging market * Factors affecting

quality and cost * Factors affecting the usability of complex documents in image form *

++

Donald WATERS, head of the Systems Office, Yale University Library, reported on the progress of a master plan for a project at Yale to convert microfilm to digital imagery, Project Open Book (POB). Stating that POB was in an advanced stage of planning, WATERS detailed, in particular, the process of selecting a vendor partner and several key issues under discussion as Yale prepares to move into the project itself. He commented first on the vision that serves as the context of POB and then described its purpose and scope.

WATERS sees the library of the future not necessarily as an electronic library but as a place that generates, preserves, and improves for its clients ready access to both intellectual and physical recorded knowledge. Electronic tools must find a place in the library in the context of this vision. Several roles for electronic tools include serving as: indirect sources of electronic knowledge or as "finding" aids (the on-line catalogues, the article-level indices, registers for documents and archives); direct sources of recorded knowledge; full-text images; and various kinds of compound sources of recorded knowledge (the so-called compound documents of Hypertext, mixed text and image, mixed-text image format, and multimedia).

POB is looking particularly at images and an image library, the uses to

which images will be put (e.g., storage, printing, browsing, and then use as input for other processes), OCR as a subsequent process to image capture, or creating an image library, and also possibly generating microfilm.

While input will come from a variety of sources, POB is considering especially input from preservation microfilm. A possible outcome is that the film and paper which provide the input for the image library eventually may go off into remote storage, and that the image library may be the primary access tool.

The purpose and scope of POB focus on imaging. Though related to CXP, POB has two features which distinguish it: 1) scale--conversion of 10,000 volumes into digital image form; and 2) source--conversion from microfilm. Given these features, several key working hypotheses guide POB, including: 1) Since POB is using microfilm, it is not concerned with the image library as a preservation medium. 2) Digital imagery can improve access to recorded knowledge through printing and network distribution at a modest incremental cost of microfilm. 3) Capturing and storing documents in a digital image form is necessary to further improvements in access. (POB distinguishes between the imaging, digitizing process and OCR, which at this stage it does not plan to perform.)

Currently in its first or organizational phase, POB found that it could use a vendor selection process to facilitate a good deal of the organizational work (e.g., creating a project team and advisory board,

confirming the validity of the plan, establishing the cost of the project and a budget, selecting the materials to convert, and then raising the necessary funds).

POB developed numerous selection criteria, including: a firm committed to image-document management, the ability to serve as systems integrator in a large-scale project over several years, interest in developing the requisite software as a standard rather than a custom product, and a willingness to invest substantial resources in the project itself.

Two vendors, DEC and Xerox, were selected as finalists in October 1991, and with the support of the Commission on Preservation and Access, each was commissioned to generate a detailed requirements analysis for the project and then to submit a formal proposal for the completion of the project, which included a budget and costs. The terms were that POB would pay the loser. The results for Yale of involving a vendor included: broad involvement of Yale staff across the board at a relatively low cost, which may have long-term significance in carrying out the project (twenty-five to thirty university people are engaged in POB); better understanding of the factors that affect corporate response to markets for imaging products; a competitive proposal; and a more sophisticated view of the imaging markets.

The most important factor that distinguished the vendors under consideration was their identification with the customer. The size and internal complexity of the company also was an important factor. POB was

looking at large companies that had substantial resources. In the end, the process generated for Yale two competitive proposals, with Xerox's the clear winner. WATERS then described the components of the proposal, the design principles, and some of the costs estimated for the process.

Components are essentially four: a conversion subsystem, a network-accessible storage subsystem for 10,000 books (and POB expects 200 to 600 dpi storage), browsing stations distributed on the campus network, and network access to the image printers.

Among the design principles, POB wanted conversion at the highest possible resolution. Assuming TIFF files, TIFF files with Group 4 compression, TCP/IP, and ethernet network on campus, POB wanted a client-server approach with image documents distributed to the workstations and made accessible through native workstation interfaces such as Windows. POB also insisted on a phased approach to implementation: 1) a stand-alone, single-user, low-cost entry into the business with a workstation focused on conversion and allowing POB to explore user access; 2) movement into a higher-volume conversion with network-accessible storage and multiple access stations; and 3) a high-volume conversion, full-capacity storage, and multiple browsing stations distributed throughout the campus.

The costs proposed for start-up assumed the existence of the Yale network and its two DocuTech image printers. Other start-up costs are estimated at $1 million over the three phases. At the end of the project, the annual

operating costs estimated primarily for the software and hardware proposed come to about $60,000, but these exclude costs for labor needed in the conversion process, network and printer usage, and facilities management.

Finally, the selection process produced for Yale a more sophisticated view of the imaging markets: the management of complex documents in image form is not a preservation problem, not a library problem, but a general problem in a broad, general industry. Preservation materials are useful for developing that market because of the qualities of the material. For example, much of it is out of copyright. The resolution of key issues such as the quality of scanning and image browsing also will affect development of that market.

The technology is readily available but changing rapidly. In this context of rapid change, several factors affect quality and cost, to which POB intends to pay particular attention, for example, the various levels of resolution that can be achieved. POB believes it can bring resolution up to 600 dpi, but an interpolation process from 400 to 600 is more likely. The variation quality in microfilm will prove to be a highly important factor. POB may reexamine the standards used to film in the first place by looking at this process as a follow-on to microfilming.

Other important factors include: the techniques available to the operator for handling material, the ways of integrating quality control into the digitizing work flow, and a work flow that includes indexing and storage. POB's requirement was to be able to deal with quality control

at the point of scanning. Thus, thanks to Xerox, POB anticipates having a mechanism which will allow it not only to scan in batch form, but to review the material as it goes through the scanner and control quality from the outset.

The standards for measuring quality and costs depend greatly on the uses of the material, including subsequent OCR, storage, printing, and browsing. But especially at issue for POB is the facility for browsing. This facility, WATERS said, is perhaps the weakest aspect of imaging technology and the most in need of development.

A variety of factors affect the usability of complex documents in image form, among them: 1) the ability of the system to handle the full range of document types, not just monographs but serials, multi-part monographs, and manuscripts; 2) the location of the database of record for bibliographic information about the image document, which POB wants to enter once and in the most useful place, the on-line catalog; 3) a document identifier for referencing the bibliographic information in one place and the images in another; 4) the technique for making the basic internal structure of the document accessible to the reader; and finally, 5) the physical presentation on the CRT of those documents. POB is ready to complete this phase now. One last decision involves deciding which material to scan.

******

++

DISCUSSION * TIFF files constitute de facto standard * NARA's experience with image conversion software and text conversion * RFC 1314 * Considerable flux concerning available hardware and software solutions * NAL through-put rate during scanning * Window management questions *

++

In the question-and-answer period that followed WATERS's presentation, the following points emerged:

* ZIDAR's statement about using TIFF files as a standard meant de facto standard. This is what most people use and typically exchange with other groups, across platforms, or even occasionally across display software.

* HOLMES commented on the unsuccessful experience of NARA in attempting to run image-conversion software or to exchange between applications: What are supposedly TIFF files go into other software that is supposed to be able to accept TIFF but cannot recognize the format and cannot deal with it, and thus renders the exchange useless. Re text conversion, he noted the different recognition rates obtained by substituting the make and model of scanners in NARA's recent test of an "intelligent" character-recognition product for a new company. In the selection of hardware and software, HOLMES argued, software no longer constitutes the overriding factor it did until about a year ago; rather it is perhaps important to

look at both now.

* Danny Cohen and Alan Katz of the University of Southern California Information Sciences Institute began circulating as an Internet RFC (RFC 1314) about a month ago a standard for a TIFF interchange format for Internet distribution of monochrome bit-mapped images, which LYNCH said he believed would be used as a de facto standard.

* FLEISCHHAUER's impression from hearing these reports and thinking about AM's experience was that there is considerable flux concerning available hardware and software solutions. HOOTON agreed and commented at the same time on ZIDAR's statement that the equipment employed affects the results produced. One cannot draw a complete conclusion by saying it is difficult or impossible to perform OCR from scanning microfilm, for example, with that device, that set of parameters, and system requirements, because numerous other people are accomplishing just that, using other components, perhaps. HOOTON opined that both the hardware and the software were highly important. Most of the problems discussed today have been solved in numerous different ways by other people. Though it is good to be cognizant of various experiences, this is not to say that it will always be thus.

* At NAL, the through-put rate of the scanning process for paper, page by page, performing OCR, ranges from 300 to 600 pages per day; not performing OCR is considerably faster, although how much faster

is not known. This is for scanning from bound books, which is much slower.

* WATERS commented on window management questions: DEC proposed an X-Windows solution which was problematical for two reasons. One was POB's requirement to be able to manipulate images on the workstation and bring them down to the workstation itself and the other was network usage.

******

++

THOMA * Illustration of deficiencies in scanning and storage process * Image quality in this process * Different costs entailed by better image quality * Techniques for overcoming various de-ficiencies: fixed thresholding, dynamic thresholding, dithering, image merge * Page edge effects *

++

George THOMA, chief, Communications Engineering Branch, National Library of Medicine (NLM), illustrated several of the deficiencies discussed by the previous speakers. He introduced the topic of special problems by noting the advantages of electronic imaging. For example, it is regenerable because it is a coded file, and real-time quality control is possible with electronic capture, whereas in photographic capture it is not.

One of the difficulties discussed in the scanning and storage process was image quality which, without belaboring the obvious, means different things for maps, medical X-rays, or broadcast television. In the case of documents, THOMA said, image quality boils down to legibility of the textual parts, and fidelity in the case of gray or color photo print-type material. Legibility boils down to scan density, the standard in most cases being 300 dpi. Increasing the resolution with scanners that perform 600 or 1200 dpi, however, comes at a cost.

Better image quality entails at least four different kinds of costs: 1) equipment costs, because the CCD (i.e., charge-couple device) with greater number of elements costs more; 2) time costs that translate to the actual capture costs, because manual labor is involved (the time is also dependent on the fact that more data has to be moved around in the machine in the scanning or network devices that perform the scanning as well as the storage); 3) media costs, because at high resolutions larger files have to be stored; and 4) transmission costs, because there is just more data to be transmitted.

But while resolution takes care of the issue of legibility in image quality, other deficiencies have to do with contrast and elements on the page scanned or the image that needed to be removed or clarified. Thus, THOMA proceeded to illustrate various deficiencies, how they are manifested, and several techniques to overcome them.

Fixed thresholding was the first technique described, suitable for black-and-white text, when the contrast does not vary over the page. One can have many different threshold levels in scanning devices. Thus, THOMA offered an example of extremely poor contrast, which resulted from the fact that the stock was a heavy red. This is the sort of image that when microfilmed fails to provide any legibility whatsoever. Fixed thresholding is the way to change the black-to-red contrast to the desired black-to-white contrast.

Other examples included material that had been browned or yellowed by age. This was also a case of contrast deficiency, and correction was done by fixed thresholding. A final example boils down to the same thing, slight variability, but it is not significant. Fixed thresholding solves this problem as well. The microfilm equivalent is certainly legible, but it comes with dark areas. Though THOMA did not have a slide of the microfilm in this case, he did show the reproduced electronic image.

When one has variable contrast over a page or the lighting over the page area varies, especially in the case where a bound volume has light shining on it, the image must be processed by a dynamic thresholding scheme. One scheme, dynamic averaging, allows the threshold level not to be fixed but to be recomputed for every pixel from the neighboring characteristics. The neighbors of a pixel determine where the threshold should be set for that pixel.

THOMA showed an example of a page that had been made deficient by a

variety of techniques, including a burn mark, coffee stains, and a yellow marker. Application of a fixed-thresholding scheme, THOMA argued, might take care of several deficiencies on the page but not all of them. Performing the calculation for a dynamic threshold setting, however, removes most of the deficiencies so that at least the text is legible.

Another problem is representing a gray level with black-and-white pixels by a process known as dithering or electronic screening. But dithering does not provide good image quality for pure black-and-white textual material. THOMA illustrated this point with examples. Although its suitability for photoprint is the reason for electronic screening or dithering, it cannot be used for every compound image. In the document that was distributed by CXP, THOMA noticed that the dithered image of the IEEE test chart evinced some deterioration in the text. He presented an extreme example of deterioration in the text in which compounded documents had to be set right by other techniques. The technique illustrated by the present example was an image merge in which the page is scanned twice and the settings go from fixed threshold to the dithering matrix; the resulting images are merged to give the best results with each technique.

THOMA illustrated how dithering is also used in nonphotographic or nonprint materials with an example of a grayish page from a medical text, which was reproduced to show all of the gray that appeared in the original. Dithering provided a reproduction of all the gray in the original of another example from the same text.

THOMA finally illustrated the problem of bordering, or page-edge, effects. Books and bound volumes that are placed on a photocopy machine or a scanner produce page-edge effects that are undesirable for two reasons: 1) the aesthetics of the image; after all, if the image is to be preserved, one does not necessarily want to keep all of its deficiencies; 2) compression (with the bordering problem THOMA illustrated, the compression ratio deteriorated tremendously). One way to eliminate this more serious problem is to have the operator at the point of scanning window the part of the image that is desirable and automatically turn all of the pixels out of that picture to white.

******

++

FLEISCHHAUER * AM's experience with scanning bound materials * Dithering *

++

Carl FLEISCHHAUER, coordinator, American Memory, Library of Congress, reported AM's experience with scanning bound materials, which he likened to the problems involved in using photocopying machines. Very few devices in the industry offer book-edge scanning, let alone book cradles. The problem may be unsolvable, FLEISCHHAUER said, because a large enough market does not exist for a preservation-quality scanner. AM is using a Kurzweil scanner, which is a book-edge scanner now sold by Xerox.

Devoting the remainder of his brief presentation to dithering, FLEISCHHAUER related AM's experience with a contractor who was using unsophisticated equipment and software to reduce moire patterns from printed halftones. AM took the same image and used the dithering algorithm that forms part of the same Kurzweil Xerox scanner; it disguised moire patterns much more effectively.

FLEISCHHAUER also observed that dithering produces a binary file which is useful for numerous purposes, for example, printing it on a laser printer without having to "re-halftone" it. But it tends to defeat efficient compression, because the very thing that dithers to reduce moire patterns also tends to work against compression schemes. AM thought the difference in image quality was worth it.

******

++

DISCUSSION * Relative use as a criterion for POB's selection of books to be converted into digital form *

++

During the discussion period, WATERS noted that one of the criteria for selecting books among the 10,000 to be converted into digital image form would be how much relative use they would receive--a subject still

requiring evaluation. The challenge will be to understand whether coherent bodies of material will increase usage or whether POB should seek material that is being used, scan that, and make it more accessible. POB might decide to digitize materials that are already heavily used, in order to make them more accessible and decrease wear on them. Another approach would be to provide a large body of intellectually coherent material that may be used more in digital form than it is currently used in microfilm. POB would seek material that was out of copyright.

******

++

BARONAS * Origin and scope of AIIM * Types of documents produced in AIIM's standards program * Domain of AIIM's standardization work * AIIM's structure * TC 171 and MS23 * Electronic image management standards * Categories of EIM standardization where AIIM standards are being developed *

++

Jean BARONAS, senior manager, Department of Standards and Technology, Association for Information and Image Management (AIIM), described the not-for-profit association and the national and international programs for standardization in which AIIM is active.

Accredited for twenty-five years as the nation's standards development organization for document image management, AIIM began life in a library

community developing microfilm standards. Today the association maintains both its library and business-image management standardization activities--and has moved into electronic image-management standardization (EIM).

BARONAS defined the program's scope. AIIM deals with: 1) the terminology of standards and of the technology it uses; 2) methods of measurement for the systems, as well as quality; 3) methodologies for users to evaluate and measure quality; 4) the features of apparatus used to manage and edit images; and 5) the procedures used to manage images.

BARONAS noted that three types of documents are produced in the AIIM standards program: the first two, accredited by the American National Standards Institute (ANSI), are standards and standard recommended practices. Recommended practices differ from standards in that they contain more tutorial information. A technical report is not an ANSI standard. Because AIIM's policies and procedures for developing standards are approved by ANSI, its standards are labeled ANSI/AIIM, followed by the number and title of the standard.

BARONAS then illustrated the domain of AIIM's standardization work. For example, AIIM is the administrator of the U.S. Technical Advisory Group (TAG) to the International Standards Organization's (ISO) technical committee, TC l7l Micrographics and Optical Memories for Document and Image Recording, Storage, and Use. AIIM officially works through ANSI in the international standardization process.

BARONAS described AIIM's structure, including its board of directors, its standards board of twelve individuals active in the image-management industry, its strategic planning and legal admissibility task forces, and its National Standards Council, which is comprised of the members of a number of organizations who vote on every AIIM standard before it is published. BARONAS pointed out that AIIM's liaisons deal with numerous other standards developers, including the optical disk community, office and publishing systems, image-codes-and-character set committees, and the National Information Standards Organization (NISO).

BARONAS illustrated the procedures of TC l7l, which covers all aspects of image management. When AIIM's national program has conceptualized a new project, it is usually submitted to the international level, so that the member countries of TC l7l can simultaneously work on the development of the standard or the technical report. BARONAS also illustrated a classic microfilm standard, MS23, which deals with numerous imaging concepts that apply to electronic imaging. Originally developed in the l970s, revised in the l980s, and revised again in l991, this standard is scheduled for another revision. MS23 is an active standard whereby users may propose new density ranges and new methods of evaluating film images in the standard's revision.

BARONAS detailed several electronic image-management standards, for instance, ANSI/AIIM MS44, a quality-control guideline for scanning 8.5" by 11" black-and-white office documents. This standard is used with the

IEEE fax image--a continuous tone photographic image with gray scales, text, and several continuous tone pictures--and AIIM test target number 2, a representative document used in office document management.

BARONAS next outlined the four categories of EIM standardization in which AIIM standards are being developed: transfer and retrieval, evaluation, optical disc and document scanning applications, and design and conversion of documents. She detailed several of the main projects of each: 1) in the category of image transfer and retrieval, a bi-level image transfer format, ANSI/AIIM MS53, which is a proposed standard that describes a file header for image transfer between unlike systems when the images are compressed using G3 and G4 compression; 2) the category of image evaluation, which includes the AIIM-proposed TR26 tutorial on image resolution (this technical report will treat the differences and similarities between classical or photographic and electronic imaging); 3) design and conversion, which includes a proposed technical report called "Forms Design Optimization for EIM" (this report considers how general-purpose business forms can be best designed so that scanning is optimized; reprographic characteristics such as type, rules, background, tint, and color will likewise be treated in the technical report); 4) disk and document scanning applications includes a project a) on planning platters and disk management, b) on generating an application profile for EIM when images are stored and distributed on CD-ROM, and c) on evaluating SCSI2, and how a common command set can be generated for SCSI2 so that document scanners are more easily integrated. (ANSI/AIIM MS53 will also apply to compressed images.)

******

++

BATTIN * The implications of standards for preservation * A major obstacle to successful cooperation * A hindrance to access in the digital environment * Standards a double-edged sword for those concerned with the preservation of the human record * Near-term prognosis for reliable archival standards * Preservation concerns for electronic media * Need for reconceptualizing our preservation principles * Standards in the real world and the politics of reproduction * Need to redefine the concept of archival and to begin to think in terms of life cycles * Cooperation and the La Guardia Eight * Concerns generated by discussions on the problems of preserving text and image * General principles to be adopted in a world without standards *

++

Patricia BATTIN, president, the Commission on Preservation and Access (CPA), addressed the implications of standards for preservation. She listed several areas where the library profession and the analog world of the printed book had made enormous contributions over the past hundred years--for example, in bibliographic formats, binding standards, and, most important, in determining what constitutes longevity or archival quality.

Although standards have lightened the preservation burden through the development of national and international collaborative programs, nevertheless, a pervasive mistrust of other people's standards remains a

major obstacle to successful cooperation, BATTIN said.

The zeal to achieve perfection, regardless of the cost, has hindered rather than facilitated access in some instances, and in the digital environment, where no real standards exist, has brought an ironically just reward.

BATTIN argued that standards are a double-edged sword for those concerned with the preservation of the human record, that is, the provision of access to recorded knowledge in a multitude of media as far into the future as possible. Standards are essential to facilitate interconnectivity and access, but, BATTIN said, as LYNCH pointed out yesterday, if set too soon they can hinder creativity, expansion of capability, and the broadening of access. The characteristics of standards for digital imagery differ radically from those for analog imagery. And the nature of digital technology implies continuing volatility and change. To reiterate, precipitous standard-setting can inhibit creativity, but delayed standard-setting results in chaos.

Since in BATTIN'S opinion the near-term prognosis for reliable archival standards, as defined by librarians in the analog world, is poor, two alternatives remain: standing pat with the old technology, or reconceptualizing.

Preservation concerns for electronic media fall into two general domains. One is the continuing assurance of access to knowledge originally

generated, stored, disseminated, and used in electronic form. This domain contains several subdivisions, including 1) the closed, proprietary systems discussed the previous day, bundled information such as electronic journals and government agency records, and electronically produced or captured raw data; and 2) the application of digital technologies to the reformatting of materials originally published on a deteriorating analog medium such as acid paper or videotape.

The preservation of electronic media requires a reconceptualizing of our preservation principles during a volatile, standardless transition which may last far longer than any of us envision today. BATTIN urged the necessity of shifting focus from assessing, measuring, and setting standards for the permanence of the medium to the concept of managing continuing access to information stored on a variety of media and requiring a variety of ever-changing hardware and software for access--a fundamental shift for the library profession.

BATTIN offered a primer on how to move forward with reasonable confidence in a world without standards. Her comments fell roughly into two sections: 1) standards in the real world and 2) the politics of reproduction.

In regard to real-world standards, BATTIN argued the need to redefine the concept of archive and to begin to think in terms of life cycles. In the past, the naive assumption that paper would last forever produced a cavalier attitude toward life cycles. The transient nature of the electronic media has compelled people to recognize and accept upfront the

concept of life cycles in place of permanency.

Digital standards have to be developed and set in a cooperative context to ensure efficient exchange of information. Moreover, during this transition period, greater flexibility concerning how concepts such as backup copies and archival copies in the CXP are defined is necessary, or the opportunity to move forward will be lost.

In terms of cooperation, particularly in the university setting, BATTIN also argued the need to avoid going off in a hundred different directions. The CPA has catalyzed a small group of universities called the La Guardia Eight--because La Guardia Airport is where meetings take place--Harvard, Yale, Cornell, Princeton, Penn State, Tennessee, Stanford, and USC, to develop a digital preservation consortium to look at all these issues and develop de facto standards as we move along, instead of waiting for something that is officially blessed. Continuing to apply analog values and definitions of standards to the digital environment, BATTIN said, will effectively lead to forfeiture of the benefits of digital technology to research and scholarship.

Under the second rubric, the politics of reproduction, BATTIN reiterated an oft-made argument concerning the electronic library, namely, that it is more difficult to transform than to create, and nowhere is that belief expressed more dramatically than in the conversion of brittle books to new media. Preserving information published in electronic media involves making sure the information remains accessible and that digital

information is not lost through reproduction. In the analog world of photocopies and microfilm, the issue of fidelity to the original becomes paramount, as do issues of "Whose fidelity?" and "Whose original?"

BATTIN elaborated these arguments with a few examples from a recent study conducted by the CPA on the problems of preserving text and image. Discussions with scholars, librarians, and curators in a variety of disciplines dependent on text and image generated a variety of concerns, for example: 1) Copy what is, not what the technology is capable of. This is very important for the history of ideas. Scholars wish to know what the author saw and worked from. And make available at the workstation the opportunity to erase all the defects and enhance the presentation. 2) The fidelity of reproduction--what is good enough, what can we afford, and the difference it makes--issues of subjective versus objective resolution. 3) The differences between primary and secondary users. Restricting the definition of primary user to the one in whose discipline the material has been published runs one headlong into the reality that these printed books have had a host of other users from a host of other disciplines, who not only were looking for very different things, but who also shared values very different from those of the primary user. 4) The relationship of the standard of reproduction to new capabilities of scholarship--the browsing standard versus an archival standard. How good must the archival standard be? Can a distinction be drawn between potential users in setting standards for reproduction? Archival storage, use copies, browsing copies--ought an attempt to set standards even be made? 5) Finally, costs. How much are we prepared to pay to capture absolute fidelity? What are the trade-offs between vastly

enhanced access, degrees of fidelity, and costs?

These standards, BATTIN concluded, serve to complicate further the reproduction process, and add to the long list of technical standards that are necessary to ensure widespread access. Ways to articulate and analyze the costs that are attached to the different levels of standards must be found.

Given the chaos concerning standards, which promises to linger for the foreseeable future, BATTIN urged adoption of the following general principles:

* Strive to understand the changing information requirements of scholarly disciplines as more and more technology is integrated into the process of research and scholarly communication in order to meet future scholarly needs, not to build for the past. Capture deteriorating information at the highest affordable resolution, even though the dissemination and display technologies will lag.

* Develop cooperative mechanisms to foster agreement on protocols for document structure and other interchange mechanisms necessary for widespread dissemination and use before official standards are set.

* Accept that, in a transition period, de facto standards will have

to be developed.

* Capture information in a way that keeps all options open and
provides for total convertibility: OCR, scanning of microfilm,
producing microfilm from scanned documents, etc.

* Work closely with the generators of information and the builders
of networks and databases to ensure that continuing accessibility is
a primary concern from the beginning.

* Piggyback on standards under development for the broad market, and
avoid library-specific standards; work with the vendors, in order to
take advantage of that which is being standardized for the rest of
the world.

* Concentrate efforts on managing permanence in the digital world,
rather than perfecting the longevity of a particular medium.

******

++

DISCUSSION * Additional comments on TIFF *

++

During the brief discussion period that followed BATTIN's presentation, BARONAS explained that TIFF was not developed in collaboration with or under the auspices of AIIM. TIFF is a company product, not a standard, is owned by two corporations, and is always changing. BARONAS also observed that ANSI/AIIM MS53, a bi-level image file transfer format that allows unlike systems to exchange images, is compatible with TIFF as well as with DEC's architecture and IBM's MODCA/IOCA.

******

++

HOOTON * Several questions to be considered in discussing text conversion

*

++

HOOTON introduced the final topic, text conversion, by noting that it is becoming an increasingly important part of the imaging business. Many people now realize that it enhances their system to be able to have more and more character data as part of their imaging system. Re the issue of OCR versus rekeying, HOOTON posed several questions: How does one get text into computer-readable form? Does one use automated processes? Does one attempt to eliminate the use of operators where possible? Standards for accuracy, he said, are extremely important: it makes a major difference in cost and time whether one sets as a standard 98.5 percent acceptance or 99.5 percent. He mentioned outsourcing as a possibility for converting text. Finally, what one does with the image

to prepare it for the recognition process is also important, he said, because such preparation changes how recognition is viewed, as well as facilitates recognition itself.

******

++

LESK * Roles of participants in CORE * Data flow * The scanning process * The image interface * Results of experiments involving the use of electronic resources and traditional paper copies * Testing the issue of serendipity * Conclusions *

++

Michael LESK, executive director, Computer Science Research, Bell Communications Research, Inc. (Bellcore), discussed the Chemical Online Retrieval Experiment (CORE), a cooperative project involving Cornell University, OCLC, Bellcore, and the American Chemical Society (ACS).

LESK spoke on 1) how the scanning was performed, including the unusual feature of page segmentation, and 2) the use made of the text and the image in experiments.

Working with the chemistry journals (because ACS has been saving its typesetting tapes since the mid-1970s and thus has a significant back-run of the most important chemistry journals in the United States), CORE is

attempting to create an automated chemical library. Approximately a quarter of the pages by square inch are made up of images of quasi-pictorial material; dealing with the graphic components of the pages is extremely important. LESK described the roles of participants in CORE: 1) ACS provides copyright permission, journals on paper, journals on microfilm, and some of the definitions of the files; 2) at Bellcore, LESK chiefly performs the data preparation, while Dennis Egan performs experiments on the users of chemical abstracts, and supplies the indexing and numerous magnetic tapes; 3) Cornell provides the site of the experiment; 4) OCLC develops retrieval software and other user interfaces. Various manufacturers and publishers have furnished other help.

Concerning data flow, Bellcore receives microfilm and paper from ACS; the microfilm is scanned by outside vendors, while the paper is scanned inhouse on an Improvision scanner, twenty pages per minute at 300 dpi, which provides sufficient quality for all practical uses. LESK would prefer to have more gray level, because one of the ACS journals prints on some colored pages, which creates a problem.

Bellcore performs all this scanning, creates a page-image file, and also selects from the pages the graphics, to mix with the text file (which is discussed later in the Workshop). The user is always searching the ASCII file, but she or he may see a display based on the ASCII or a display based on the images.

LESK illustrated how the program performs page analysis, and the image

interface. (The user types several words, is presented with a list--usually of the titles of articles contained in an issue--that derives from the ASCII, clicks on an icon and receives an image that mirrors an ACS page.) LESK also illustrated an alternative interface, based on text on the ASCII, the so-called SuperBook interface from Bellcore.

LESK next presented the results of an experiment conducted by Dennis Egan and involving thirty-six students at Cornell, one third of them undergraduate chemistry majors, one third senior undergraduate chemistry majors, and one third graduate chemistry students. A third of them received the paper journals, the traditional paper copies and chemical abstracts on paper. A third received image displays of the pictures of the pages, and a third received the text display with pop-up graphics.

The students were given several questions made up by some chemistry professors. The questions fell into five classes, ranging from very easy to very difficult, and included questions designed to simulate browsing as well as a traditional information retrieval-type task.

LESK furnished the following results. In the straightforward question search--the question being, what is the phosphorus oxygen bond distance and hydroxy phosphate?--the students were told that they could take fifteen minutes and, then, if they wished, give up. The students with paper took more than fifteen minutes on average, and yet most of them gave up. The students with either electronic format, text or image, received good scores in reasonable time, hardly ever had to give up, and

usually found the right answer.

In the browsing study, the students were given a list of eight topics, told to imagine that an issue of the Journal of the American Chemical Society had just appeared on their desks, and were also told to flip through it and to find topics mentioned in the issue. The average scores were about the same. (The students were told to answer yes or no about whether or not particular topics appeared.) The errors, however, were quite different. The students with paper rarely said that something appeared when it had not. But they often failed to find something actually mentioned in the issue. The computer people found numerous things, but they also frequently said that a topic was mentioned when it was not. (The reason, of course, was that they were performing word searches. They were finding that words were mentioned and they were concluding that they had accomplished their task.)

This question also contained a trick to test the issue of serendipity. The students were given another list of eight topics and instructed, without taking a second look at the journal, to recall how many of this new list of eight topics were in this particular issue. This was an attempt to see if they performed better at remembering what they were not looking for. They all performed about the same, paper or electronics, about 62 percent accurate. In short, LESK said, people were not very good when it came to serendipity, but they were no worse at it with computers than they were with paper.

(LESK gave a parenthetical illustration of the learning curve of students who used SuperBook.)

The students using the electronic systems started off worse than the ones using print, but by the third of the three sessions in the series had caught up to print. As one might expect, electronics provide a much better means of finding what one wants to read; reading speeds, once the object of the search has been found, are about the same.

Almost none of the students could perform the hard task--the analogous transformation. (It would require the expertise of organic chemists to complete.) But an interesting result was that the students using the text search performed terribly, while those using the image system did best. That the text search system is driven by text offers the explanation. Everything is focused on the text; to see the pictures, one must press on an icon. Many students found the right article containing the answer to the question, but they did not click on the icon to bring up the right figure and see it. They did not know that they had found the right place, and thus got it wrong.

The short answer demonstrated by this experiment was that in the event one does not know what to read, one needs the electronic systems; the electronic systems hold no advantage at the moment if one knows what to read, but neither do they impose a penalty.

LESK concluded by commenting that, on one hand, the image system was easy

to use. On the other hand, the text display system, which represented twenty man-years of work in programming and polishing, was not winning, because the text was not being read, just searched. The much easier system is highly competitive as well as remarkably effective for the actual chemists.

******

++

ERWAY * Most challenging aspect of working on AM * Assumptions guiding AM's approach * Testing different types of service bureaus * AM's requirement for 99.95 percent accuracy * Requirements for text-coding * Additional factors influencing AM's approach to coding * Results of AM's experience with rekeying * Other problems in dealing with service bureaus * Quality control the most time-consuming aspect of contracting out conversion * Long-term outlook uncertain *

++

To Ricky ERWAY, associate coordinator, American Memory, Library of Congress, the constant variety of conversion projects taking place simultaneously represented perhaps the most challenging aspect of working on AM. Thus, the challenge was not to find a solution for text conversion but a tool kit of solutions to apply to LC's varied collections that need to be converted. ERWAY limited her remarks to the process of converting text to machine-readable form, and the variety of LC's text collections, for example, bound volumes, microfilm, and

handwritten manuscripts.

Two assumptions have guided AM's approach, ERWAY said: 1) A desire not to perform the conversion inhouse. Because of the variety of formats and types of texts, to capitalize the equipment and have the talents and skills to operate them at LC would be extremely expensive. Further, the natural inclination to upgrade to newer and better equipment each year made it reasonable for AM to focus on what it did best and seek external conversion services. Using service bureaus also allowed AM to have several types of operations take place at the same time. 2) AM was not a technology project, but an effort to improve access to library collections. Hence, whether text was converted using OCR or rekeying mattered little to AM. What mattered were cost and accuracy of results.

AM considered different types of service bureaus and selected three to perform several small tests in order to acquire a sense of the field. The sample collections with which they worked included handwritten correspondence, typewritten manuscripts from the 1940s, and eighteenth-century printed broadsides on microfilm. On none of these samples was OCR performed; they were all rekeyed. AM had several special requirements for the three service bureaus it had engaged. For instance, any errors in the original text were to be retained. Working from bound volumes or anything that could not be sheet-fed also constituted a factor eliminating companies that would have performed OCR.

AM requires 99.95 percent accuracy, which, though it sounds high, often

means one or two errors per page. The initial batch of test samples contained several handwritten materials for which AM did not require text-coding. The results, ERWAY reported, were in all cases fairly comparable: for the most part, all three service bureaus achieved 99.95 percent accuracy. AM was satisfied with the work but surprised at the cost.

As AM began converting whole collections, it retained the requirement for 99.95 percent accuracy and added requirements for text-coding. AM needed to begin performing work more than three years ago before LC requirements for SGML applications had been established. Since AM's goal was simply to retain any of the intellectual content represented by the formatting of the document (which would be lost if one performed a straight ASCII conversion), AM used "SGML-like" codes. These codes resembled SGML tags but were used without the benefit of document-type definitions. AM found that many service bureaus were not yet SGML-proficient.

Additional factors influencing the approach AM took with respect to coding included: 1) the inability of any known microcomputer-based user-retrieval software to take advantage of SGML coding; and 2) the multiple inconsistencies in format of the older documents, which confirmed AM in its desire not to attempt to force the different formats to conform to a single document-type definition (DTD) and thus create the need for a separate DTD for each document.

The five text collections that AM has converted or is in the process of converting include a collection of eighteenth-century broadsides, a

collection of pamphlets, two typescript document collections, and a collection of 150 books.

ERWAY next reviewed the results of AM's experience with rekeying, noting again that because the bulk of AM's materials are historical, the quality of the text often does not lend itself to OCR. While non-English speakers are less likely to guess or elaborate or correct typos in the original text, they are also less able to infer what we would; they also are nearly incapable of converting handwritten text. Another disadvantage of working with overseas keyers is that they are much less likely to telephone with questions, especially on the coding, with the result that they develop their own rules as they encounter new situations.

Government contracting procedures and time frames posed a major challenge to performing the conversion. Many service bureaus are not accustomed to retaining the image, even if they perform OCR. Thus, questions of image format and storage media were somewhat novel to many of them. ERWAY also remarked other problems in dealing with service bureaus, for example, their inability to perform text conversion from the kind of microfilm that LC uses for preservation purposes.

But quality control, in ERWAY's experience, was the most time-consuming aspect of contracting out conversion. AM has been attempting to perform a 10-percent quality review, looking at either every tenth document or every tenth page to make certain that the service bureaus are maintaining

99.95 percent accuracy. But even if they are complying with the requirement for accuracy, finding errors produces a desire to correct them and, in turn, to clean up the whole collection, which defeats the purpose to some extent. Even a double entry requires a character-by-character comparison to the original to meet the accuracy requirement. LC is not accustomed to publish imperfect texts, which makes attempting to deal with the industry standard an emotionally fraught issue for AM. As was mentioned in the previous day's discussion, going from 99.95 to 99.99 percent accuracy usually doubles costs and means a third keying or another complete run-through of the text.

Although AM has learned much from its experiences with various collections and various service bureaus, ERWAY concluded pessimistically that no breakthrough has been achieved. Incremental improvements have occurred in some of the OCR technology, some of the processes, and some of the standards acceptances, which, though they may lead to somewhat lower costs, do not offer much encouragement to many people who are anxiously awaiting the day that the entire contents of LC are available on-line.

******

++

ZIDAR * Several answers to why one attempts to perform full-text conversion * Per page cost of performing OCR * Typical problems encountered during editing * Editing poor copy OCR vs. rekeying *

++

Judith ZIDAR, coordinator, National Agricultural Text Digitizing Program (NATDP), National Agricultural Library (NAL), offered several answers to the question of why one attempts to perform full-text conversion: 1) Text in an image can be read by a human but not by a computer, so of course it is not searchable and there is not much one can do with it. 2) Some material simply requires word-level access. For instance, the legal profession insists on full-text access to its material; with taxonomic or geographic material, which entails numerous names, one virtually requires word-level access. 3) Full text permits rapid browsing and searching, something that cannot be achieved in an image with today's technology. 4) Text stored as ASCII and delivered in ASCII is standardized and highly portable. 5) People just want full-text searching, even those who do not know how to do it. NAL, for the most part, is performing OCR at an actual cost per average-size page of approximately $7. NAL scans the page to create the electronic image and passes it through the OCR device.

ZIDAR next rehearsed several typical problems encountered during editing. Praising the celerity of her student workers, ZIDAR observed that editing requires approximately five to ten minutes per page, assuming that there are no large tables to audit. Confusion among the three characters I, 1, and l, constitutes perhaps the most common problem encountered. Zeroes and O's also are frequently confused. Double M's create a particular problem, even on clean pages. They are so wide in most fonts that they touch, and the system simply cannot tell where one letter ends and the other begins. Complex page formats occasionally fail to columnate properly, which entails rescanning as though one were working with a

single column, entering the ASCII, and decolumnating for better searching. With proportionally spaced text, OCR can have difficulty discerning what is a space and what are merely spaces between letters, as opposed to spaces between words, and therefore will merge text or break up words where it should not.

ZIDAR said that it can often take longer to edit a poor-copy OCR than to key it from scratch. NAL has also experimented with partial editing of text, whereby project workers go into and clean up the format, removing stray characters but not running a spell-check. NAL corrects typos in the title and authors' names, which provides a foothold for searching and browsing. Even extremely poor-quality OCR (e.g., 60-percent accuracy) can still be searched, because numerous words are correct, while the important words are probably repeated often enough that they are likely to be found correct somewhere. Librarians, however, cannot tolerate this situation, though end users seem more willing to use this text for searching, provided that NAL indicates that it is unedited. ZIDAR concluded that rekeying of text may be the best route to take, in spite of numerous problems with quality control and cost.

******

++

DISCUSSION * Modifying an image before performing OCR * NAL's costs per page *AM's costs per page and experience with Federal Prison Industries * Elements comprising NATDP's costs per page * OCR and structured markup *

Distinction between the structure of a document and its representation

when put on the screen or printed *

++

HOOTON prefaced the lengthy discussion that followed with several comments about modifying an image before one reaches the point of performing OCR. For example, in regard to an application containing a significant amount of redundant data, such as form-type data, numerous companies today are working on various kinds of form renewal, prior to going through a recognition process, by using dropout colors. Thus, acquiring access to form design or using electronic means are worth considering. HOOTON also noted that conversion usually makes or breaks one's imaging system. It is extremely important, extremely costly in terms of either capital investment or service, and determines the quality of the remainder of one's system, because it determines the character of the raw material used by the system.

Concerning the four projects undertaken by NAL, two inside and two performed by outside contractors, ZIDAR revealed that an in-house service bureau executed the first at a cost between $8 and $10 per page for everything, including building of the database. The project undertaken by the Consultative Group on International Agricultural Research (CGIAR) cost approximately $10 per page for the conversion, plus some expenses for the software and building of the database. The Acid Rain Project--a two-disk set produced by the University of Vermont, consisting of Canadian publications on acid rain--cost $6.70 per page for everything, including keying of the text, which was double keyed, scanning of the

images, and building of the database. The in-house project offered considerable ease of convenience and greater control of the process. On the other hand, the service bureaus know their job and perform it expeditiously, because they have more people.

As a useful comparison, ERWAY revealed AM's costs as follows: $0.75 cents to $0.85 cents per thousand characters, with an average page containing 2,700 characters. Requirements for coding and imaging increase the costs. Thus, conversion of the text, including the coding, costs approximately $3 per page. (This figure does not include the imaging and database-building included in the NAL costs.) AM also enjoyed a happy experience with Federal Prison Industries, which precluded the necessity of going through the request-for-proposal process to award a contract, because it is another government agency. The prisoners performed AM's rekeying just as well as other service bureaus and proved handy as well. AM shipped them the books, which they would photocopy on a book-edge scanner. They would perform the markup on photocopies, return the books as soon as they were done with them, perform the keying, and return the material to AM on WORM disks.

ZIDAR detailed the elements that constitute the previously noted cost of approximately $7 per page. Most significant is the editing, correction of errors, and spell-checkings, which though they may sound easy to perform require, in fact, a great deal of time. Reformatting text also takes a while, but a significant amount of NAL's expenses are for equipment, which was extremely expensive when purchased because it was one of the few systems on the market. The costs of equipment are being amortized over

five years but are still quite high, nearly $2,000 per month.

HOCKEY raised a general question concerning OCR and the amount of editing required (substantial in her experience) to generate the kind of structured markup necessary for manipulating the text on the computer or loading it into any retrieval system. She wondered if the speakers could extend the previous question about the cost-benefit of adding or exerting structured markup. ERWAY noted that several OCR systems retain italics, bolding, and other spatial formatting. While the material may not be in the format desired, these systems possess the ability to remove the original materials quickly from the hands of the people performing the conversion, as well as to retain that information so that users can work with it. HOCKEY rejoined that the current thinking on markup is that one should not say that something is italic or bold so much as why it is that way. To be sure, one needs to know that something was italicized, but how can one get from one to the other? One can map from the structure to the typographic representation.

FLEISCHHAUER suggested that, given the 100 million items the Library holds, it may not be possible for LC to do more than report that a thing was in italics as opposed to why it was italics, although that may be desirable in some contexts. Promising to talk a bit during the afternoon session about several experiments OCLC performed on automatic recognition of document elements, and which they hoped to extend, WEIBEL said that in fact one can recognize the major elements of a document with a fairly high degree of reliability, at least as good as OCR. STEVENS drew a useful distinction between standard, generalized markup (i.e., defining

for a document-type definition the structure of the document), and what he termed a style sheet, which had to do with italics, bolding, and other forms of emphasis. Thus, two different components are at work, one being the structure of the document itself (its logic), and the other being its representation when it is put on the screen or printed.

******

SESSION V. APPROACHES TO PREPARING ELECTRONIC TEXTS

++

HOCKEY * Text in ASCII and the representation of electronic text versus an image * The need to look at ways of using markup to assist retrieval * The need for an encoding format that will be reusable and multifunctional

++

Susan HOCKEY, director, Center for Electronic Texts in the Humanities (CETH), Rutgers and Princeton Universities, announced that one talk (WEIBEL's) was moved into this session from the morning and that David Packard was unable to attend. The session would attempt to focus more on what one can do with a text in ASCII and the representation of electronic text rather than just an image, what one can do with a computer that cannot be done with a book or an image. It would be argued that one can do much more than just read a text, and from that starting point one can use markup and methods of preparing the text to take full advantage of the capability of the computer. That would lead to a discussion of what

the European Community calls REUSABILITY, what may better be termed DURABILITY, that is, how to prepare or make a text that will last a long time and that can be used for as many applications as possible, which would lead to issues of improving intellectual access.

HOCKEY urged the need to look at ways of using markup to facilitate retrieval, not just for referencing or to help locate an item that is retrieved, but also to put markup tags in a text to help retrieve the thing sought either with linguistic tagging or interpretation. HOCKEY also argued that little advancement had occurred in the software tools currently available for retrieving and searching text. She pressed the desideratum of going beyond Boolean searches and performing more sophisticated searching, which the insertion of more markup in the text would facilitate. Thinking about electronic texts as opposed to images means considering material that will never appear in print form, or print will not be its primary form, that is, material which only appears in electronic form. HOCKEY alluded to the history and the need for markup and tagging and electronic text, which was developed through the use of computers in the humanities; as MICHELSON had observed, Father Busa had started in 1949 to prepare the first-ever text on the computer.

HOCKEY remarked several large projects, particularly in Europe, for the compilation of dictionaries, language studies, and language analysis, in which people have built up archives of text and have begun to recognize the need for an encoding format that will be reusable and multifunctional, that can be used not just to print the text, which may be assumed to be a byproduct of what one wants to do, but to structure it inside the computer so that it can be searched, built into a Hypertext system, etc.

******

++

WEIBEL * OCLC's approach to preparing electronic text: retroconversion,
keying of texts, more automated ways of developing data * Project ADAPT
and the CORE Project * Intelligent character recognition does not exist *
Advantages of SGML * Data should be free of procedural markup;
descriptive markup strongly advocated * OCLC's interface illustrated *
Storage requirements and costs for putting a lot of information on line *

++

Stuart WEIBEL, senior research scientist, Online Computer Library Center,
Inc. (OCLC), described OCLC's approach to preparing electronic text. He
argued that the electronic world into which we are moving must
accommodate not only the future but the past as well, and to some degree
even the present. Thus, starting out at one end with retroconversion and
keying of texts, one would like to move toward much more automated ways
of developing data.

For example, Project ADAPT had to do with automatically converting
document images into a structured document database with OCR text as
indexing and also a little bit of automatic formatting and tagging of
that text. The CORE project hosted by Cornell University, Bellcore,
OCLC, the American Chemical Society, and Chemical Abstracts, constitutes
WEIBEL's principal concern at the moment. This project is an example of

converting text for which one already has a machine-readable version into a format more suitable for electronic delivery and database searching. (Since Michael LESK had previously described CORE, WEIBEL would say little concerning it.) Borrowing a chemical phrase, de novo synthesis, WEIBEL cited the Online Journal of Current Clinical Trials as an example of de novo electronic publishing, that is, a form in which the primary form of the information is electronic.

Project ADAPT, then, which OCLC completed a couple of years ago and in fact is about to resume, is a model in which one takes page images either in paper or microfilm and converts them automatically to a searchable electronic database, either on-line or local. The operating assumption is that accepting some blemishes in the data, especially for retroconversion of materials, will make it possible to accomplish more. Not enough money is available to support perfect conversion.

WEIBEL related several steps taken to perform image preprocessing (processing on the image before performing optical character recognition), as well as image postprocessing. He denied the existence of intelligent character recognition and asserted that what is wanted is page recognition, which is a long way off. OCLC has experimented with merging of multiple optical character recognition systems that will reduce errors from an unacceptable rate of 5 characters out of every l,000 to an unacceptable rate of 2 characters out of every l,000, but it is not good enough. It will never be perfect.

Concerning the CORE Project, WEIBEL observed that Bellcore is taking the topography files, extracting the page images, and converting those topography files to SGML markup. LESK hands that data off to OCLC, which builds that data into a Newton database, the same system that underlies the on-line system in virtually all of the reference products at OCLC. The long-term goal is to make the systems interoperable so that not just Bellcore's system and OCLC's system can access this data, but other systems can as well, and the key to that is the Z39.50 common command language and the full-text extension. Z39.50 is fine for MARC records, but is not enough to do it for full text (that is, make full texts interoperable).

WEIBEL next outlined the critical role of SGML for a variety of purposes, for example, as noted by HOCKEY, in the world of extremely large databases, using highly structured data to perform field searches. WEIBEL argued that by building the structure of the data in (i.e., the structure of the data originally on a printed page), it becomes easy to look at a journal article even if one cannot read the characters and know where the title or author is, or what the sections of that document would be. OCLC wants to make that structure explicit in the database, because it will be important for retrieval purposes.

The second big advantage of SGML is that it gives one the ability to build structure into the database that can be used for display purposes without contaminating the data with instructions about how to format things. The distinction lies between procedural markup, which tells one where to put dots on the page, and descriptive markup, which describes

the elements of a document.

WEIBEL believes that there should be no procedural markup in the data at all, that the data should be completely unsullied by information about italics or boldness. That should be left up to the display device, whether that display device is a page printer or a screen display device. By keeping one's database free of that kind of contamination, one can make decisions down the road, for example, reorganize the data in ways that are not cramped by built-in notions of what should be italic and what should be bold. WEIBEL strongly advocated descriptive markup. As an example, he illustrated the index structure in the CORE data. With subsequent illustrated examples of markup, WEIBEL acknowledged the common complaint that SGML is hard to read in its native form, although markup decreases considerably once one gets into the body. Without the markup, however, one would not have the structure in the data. One can pass markup through a LaTeX processor and convert it relatively easily to a printed version of the document.

WEIBEL next illustrated an extremely cluttered screen dump of OCLC's system, in order to show as much as possible the inherent capability on the screen. (He noted parenthetically that he had become a supporter of X-Windows as a result of the progress of the CORE Project.) WEIBEL also illustrated the two major parts of the interface: l) a control box that allows one to generate lists of items, which resembles a small table of contents based on key words one wishes to search, and 2) a document viewer, which is a separate process in and of itself. He demonstrated how to follow links through the electronic database simply by selecting

the appropriate button and bringing them up. He also noted problems that remain to be accommodated in the interface (e.g., as pointed out by LESK, what happens when users do not click on the icon for the figure).

Given the constraints of time, WEIBEL omitted a large number of ancillary items in order to say a few words concerning storage requirements and what will be required to put a lot of things on line. Since it is extremely expensive to reconvert all of this data, especially if it is just in paper form (and even if it is in electronic form in typesetting tapes), he advocated building journals electronically from the start. In that case, if one only has text graphics and indexing (which is all that one needs with de novo electronic publishing, because there is no need to go back and look at bit-maps of pages), one can get 10,000 journals of full text, or almost 6 million pages per year. These pages can be put in approximately 135 gigabytes of storage, which is not all that much, WEIBEL said. For twenty years, something less than three terabytes would be required. WEIBEL calculated the costs of storing this information as follows: If a gigabyte costs approximately $1,000, then a terabyte costs approximately $1 million to buy in terms of hardware. One also needs a building to put it in and a staff like OCLC to handle that information. So, to support a terabyte, multiply by five, which gives $5 million per year for a supported terabyte of data.

******

++

DISCUSSION * Tapes saved by ACS are the typography files originally supporting publication of the journal * Cost of building tagged text into the database *

++

During the question-and-answer period that followed WEIBEL's presentation, these clarifications emerged. The tapes saved by the American Chemical Society are the typography files that originally supported the publication of the journal. Although they are not tagged in SGML, they are tagged in very fine detail. Every single sentence is marked, all the registry numbers, all the publications issues, dates, and volumes. No cost figures on tagging material on a per-megabyte basis were available. Because ACS's typesetting system runs from tagged text, there is no extra cost per article. It was unknown what it costs ACS to keyboard the tagged text rather than just keyboard the text in the cheapest process. In other words, since one intends to publish things and will need to build tagged text into a typography system in any case, if one does that in such a way that it can drive not only typography but an electronic system (which is what ACS intends to do--move to SGML publishing), the marginal cost is zero. The marginal cost represents the cost of building tagged text into the database, which is small.

******

++

SPERBERG-McQUEEN * Distinction between texts and computers * Implications

of recognizing that all representation is encoding * Dealing with complicated representations of text entails the need for a grammar of documents * Variety of forms of formal grammars * Text as a bit-mapped image does not represent a serious attempt to represent text in electronic form * SGML, the TEI, document-type declarations, and the reusability and longevity of data * TEI conformance explicitly allows extension or modification of the TEI tag set * Administrative background of the TEI * Several design goals for the TEI tag set * An absolutely fixed requirement of the TEI Guidelines * Challenges the TEI has attempted to face * Good texts not beyond economic feasibility * The issue of reproducibility or processability * The issue of mages as simulacra for the text redux * One's model of text determines what one's software can do with a text and has economic consequences *

++

Prior to speaking about SGML and markup, Michael SPERBERG-McQUEEN, editor, Text Encoding Initiative (TEI), University of Illinois-Chicago, first drew a distinction between texts and computers: Texts are abstract cultural and linguistic objects while computers are complicated physical devices, he said. Abstract objects cannot be placed inside physical devices; with computers one can only represent text and act upon those representations.

The recognition that all representation is encoding, SPERBERG-McQUEEN argued, leads to the recognition of two things: 1) The topic description for this session is slightly misleading, because there can be no discussion of pros and cons of text-coding unless what one means is pros and cons of working with text with computers. 2) No text can be represented in a

computer without some sort of encoding; images are one way of encoding text, ASCII is another, SGML yet another. There is no encoding without some information loss, that is, there is no perfect reproduction of a text that allows one to do away with the original. Thus, the question becomes, What is the most useful representation of text for a serious work? This depends on what kind of serious work one is talking about.

The projects demonstrated the previous day all involved highly complex information and fairly complex manipulation of the textual material. In order to use that complicated information, one has to calculate it slowly or manually and store the result. It needs to be stored, therefore, as part of one's representation of the text. Thus, one needs to store the structure in the text. To deal with complicated representations of text, one needs somehow to control the complexity of the representation of a text; that means one needs a way of finding out whether a document and an electronic representation of a document is legal or not; and that means one needs a grammar of documents.

SPERBERG-McQUEEN discussed the variety of forms of formal grammars, implicit and explicit, as applied to text, and their capabilities. He argued that these grammars correspond to different models of text that different developers have. For example, one implicit model of the text is that there is no internal structure, but just one thing after another, a few characters and then perhaps a start-title command, and then a few more characters and an end-title command. SPERBERG-McQUEEN also distinguished several kinds of text that have a sort of hierarchical structure that is not very well defined, which, typically, corresponds

to grammars that are not very well defined, as well as hierarchies that are very well defined (e.g., the Thesaurus Linguae Graecae) and extremely complicated things such as SGML, which handle strictly hierarchical data very nicely.

SPERBERG-McQUEEN conceded that one other model not illustrated on his two displays was the model of text as a bit-mapped image, an image of a page, and confessed to having been converted to a limited extent by the Workshop to the view that electronic images constitute a promising, probably superior alternative to microfilming. But he was not convinced that electronic images represent a serious attempt to represent text in electronic form. Many of their problems stem from the fact that they are not direct attempts to represent the text but attempts to represent the page, thus making them representations of representations.

In this situation of increasingly complicated textual information and the need to control that complexity in a useful way (which begs the question of the need for good textual grammars), one has the introduction of SGML. With SGML, one can develop specific document-type declarations for specific text types or, as with the TEI, attempts to generate general document-type declarations that can handle all sorts of text. The TEI is an attempt to develop formats for text representation that will ensure the kind of reusability and longevity of data discussed earlier. It offers a way to stay alive in the state of permanent technological revolution.

It has been a continuing challenge in the TEI to create document grammars that do some work in controlling the complexity of the textual object but also allowing one to represent the real text that one will find. Fundamental to the notion of the TEI is that TEI conformance allows one the ability to extend or modify the TEI tag set so that it fits the text that one is attempting to represent.

SPERBERG-McQUEEN next outlined the administrative background of the TEI. The TEI is an international project to develop and disseminate guidelines for the encoding and interchange of machine-readable text. It is sponsored by the Association for Computers in the Humanities, the Association for Computational Linguistics, and the Association for Literary and Linguistic Computing. Representatives of numerous other professional societies sit on its advisory board. The TEI has a number of affiliated projects that have provided assistance by testing drafts of the guidelines.

Among the design goals for the TEI tag set, the scheme first of all must meet the needs of research, because the TEI came out of the research community, which did not feel adequately served by existing tag sets. The tag set must be extensive as well as compatible with existing and emerging standards. In 1990, version 1.0 of the Guidelines was released (SPERBERG-McQUEEN illustrated their contents).

SPERBERG-McQUEEN noted that one problem besetting electronic text has been the lack of adequate internal or external documentation for many

existing electronic texts. The TEI guidelines as currently formulated contain few fixed requirements, but one of them is this: There must always be a document header, an in-file SGML tag that provides 1) a bibliographic description of the electronic object one is talking about (that is, who included it, when, what for, and under which title); and 2) the copy text from which it was derived, if any. If there was no copy text or if the copy text is unknown, then one states as much. Version 2.0 of the Guidelines was scheduled to be completed in fall 1992 and a revised third version is to be presented to the TEI advisory board for its endorsement this coming winter. The TEI itself exists to provide a markup language, not a marked-up text.

Among the challenges the TEI has attempted to face is the need for a markup language that will work for existing projects, that is, handle the level of markup that people are using now to tag only chapter, section, and paragraph divisions and not much else. At the same time, such a language also will be able to scale up gracefully to handle the highly detailed markup which many people foresee as the future destination of much electronic text, and which is not the future destination but the present home of numerous electronic texts in specialized areas.

SPERBERG-McQUEEN dismissed the lowest-common-denominator approach as unable to support the kind of applications that draw people who have never been in the public library regularly before, and make them come back. He advocated more interesting text and more intelligent text. Asserting that it is not beyond economic feasibility to have good texts, SPERBERG-McQUEEN noted that the TEI Guidelines listing 200-odd tags

contains tags that one is expected to enter every time the relevant textual feature occurs. It contains all the tags that people need now, and it is not expected that everyone will tag things in the same way.

The question of how people will tag the text is in large part a function of their reaction to what SPERBERG-McQUEEN termed the issue of reproducibility. What one needs to be able to reproduce are the things one wants to work with. Perhaps a more useful concept than that of reproducibility or recoverability is that of processability, that is, what can one get from an electronic text without reading it again in the original. He illustrated this contention with a page from Jan Comenius's bilingual Introduction to Latin.

SPERBERG-McQUEEN returned at length to the issue of images as simulacra for the text, in order to reiterate his belief that in the long run more than images of pages of particular editions of the text are needed, because just as second-generation photocopies and second-generation microfilm degenerate, so second-generation representations tend to degenerate, and one tends to overstress some relatively trivial aspects of the text such as its layout on the page, which is not always significant, despite what the text critics might say, and slight other pieces of information such as the very important lexical ties between the English and Latin versions of Comenius's bilingual text, for example. Moreover, in many crucial respects it is easy to fool oneself concerning what a scanned image of the text will accomplish. For example, in order to study the transmission of texts, information concerning the text carrier is necessary, which scanned images simply do not always handle.

Further, even the high-quality materials being produced at Cornell use much of the information that one would need if studying those books as physical objects. It is a choice that has been made. It is an arguably justifiable choice, but one does not know what color those pen strokes in the margin are or whether there was a stain on the page, because it has been filtered out. One does not know whether there were rips in the page because they do not show up, and on a couple of the marginal marks one loses half of the mark because the pen is very light and the scanner failed to pick it up, and so what is clearly a checkmark in the margin of the original becomes a little scoop in the margin of the facsimile. Standard problems for facsimile editions, not new to electronics, but also true of light-lens photography, and are remarked here because it is important that we not fool ourselves that even if we produce a very nice image of this page with good contrast, we are not replacing the manuscript any more than microfilm has replaced the manuscript.

The TEI comes from the research community, where its first allegiance lies, but it is not just an academic exercise. It has relevance far beyond those who spend all of their time studying text, because one's model of text determines what one's software can do with a text. Good models lead to good software. Bad models lead to bad software. That has economic consequences, and it is these economic consequences that have led the European Community to help support the TEI, and that will lead, SPERBERG-McQUEEN hoped, some software vendors to realize that if they provide software with a better model of the text they can make a killing.

******

++

DISCUSSION * Implications of different DTDs and tag sets * ODA versus SGML *

++

During the discussion that followed, several additional points were made. Neither AAP (i.e., Association of American Publishers) nor CALS (i.e., Computer-aided Acquisition and Logistics Support) has a document-type definition for ancient Greek drama, although the TEI will be able to handle that. Given this state of affairs and assuming that the technical-journal producers and the commercial vendors decide to use the other two types, then an institution like the Library of Congress, which might receive all of their publications, would have to be able to handle three different types of document definitions and tag sets and be able to distinguish among them.

Office Document Architecture (ODA) has some advantages that flow from its tight focus on office documents and clear directions for implementation. Much of the ODA standard is easier to read and clearer at first reading than the SGML standard, which is extremely general. What that means is that if one wants to use graphics in TIFF and ODA, one is stuck, because ODA defines graphics formats while TIFF does not, whereas SGML says the world is not waiting for this work group to create another graphics format. What is needed is an ability to use whatever graphics format one wants.

The TEI provides a socket that allows one to connect the SGML document to

the graphics. The notation that the graphics are in is clearly a choice that one needs to make based on her or his environment, and that is one advantage. SGML is less megalomaniacal in attempting to define formats for all kinds of information, though more megalomaniacal in attempting to cover all sorts of documents. The other advantage is that the model of text represented by SGML is simply an order of magnitude richer and more flexible than the model of text offered by ODA. Both offer hierarchical structures, but SGML recognizes that the hierarchical model of the text that one is looking at may not have been in the minds of the designers, whereas ODA does not.

ODA is not really aiming for the kind of document that the TEI wants to encompass. The TEI can handle the kind of material ODA has, as well as a significantly broader range of material. ODA seems to be very much focused on office documents, which is what it started out being called-- office document architecture.

******

++

CALALUCA * Text-encoding from a publisher's perspective * Responsibilities of a publisher * Reproduction of Migne's Latin series whole and complete with SGML tags based on perceived need and expected use * Particular decisions arising from the general decision to produce and publish PLD *

++

The final speaker in this session, Eric CALALUCA, vice president, Chadwyck-Healey, Inc., spoke from the perspective of a publisher re text-encoding, rather than as one qualified to discuss methods of encoding data, and observed that the presenters sitting in the room, whether they had chosen to or not, were acting as publishers: making choices, gathering data, gathering information, and making assessments. CALALUCA offered the hard-won conviction that in publishing very large text files (such as PLD), one cannot avoid making personal judgments of appropriateness and structure.

In CALALUCA's view, encoding decisions stem from prior judgments. Two notions have become axioms for him in the consideration of future sources for electronic publication: 1) electronic text publishing is as personal as any other kind of publishing, and questions of if and how to encode the data are simply a consequence of that prior decision; 2) all personal decisions are open to criticism, which is unavoidable.

CALALUCA rehearsed his role as a publisher or, better, as an intermediary between what is viewed as a sound idea and the people who would make use of it. Finding the specialist to advise in this process is the core of that function. The publisher must monitor and hug the fine line between giving users what they want and suggesting what they might need. One responsibility of a publisher is to represent the desires of scholars and research librarians as opposed to bullheadedly forcing them into areas they would not choose to enter.

CALALUCA likened the questions being raised today about data structure and standards to the decisions faced by the Abbe Migne himself during production of the Patrologia series in the mid-nineteenth century. Chadwyck-Healey's decision to reproduce Migne's Latin series whole and complete with SGML tags was also based upon a perceived need and an expected use. In the same way that Migne's work came to be far more than a simple handbook for clerics, PLD is already far more than a database for theologians. It is a bedrock source for the study of Western civilization, CALALUCA asserted.

In regard to the decision to produce and publish PLD, the editorial board offered direct judgments on the question of appropriateness of these texts for conversion, their encoding and their distribution, and concluded that the best possible project was one that avoided overt intrusions or exclusions in so important a resource. Thus, the general decision to transmit the original collection as clearly as possible with the widest possible avenues for use led to other decisions: 1) To encode the data or not, SGML or not, TEI or not. Again, the expected user community asserted the need for normative tagging structures of important humanities texts, and the TEI seemed the most appropriate structure for that purpose. Research librarians, who are trained to view the larger impact of electronic text sources on 80 or 90 or 100 doctoral disciplines, loudly approved the decision to include tagging. They see what is coming better than the specialist who is completely focused on one edition of Ambrose's De Anima, and they also understand that the potential uses exceed present expectations. 2) What will be tagged and

what will not. Once again, the board realized that one must tag the obvious. But in no way should one attempt to identify through encoding schemes every single discrete area of a text that might someday be searched. That was another decision. Searching by a column number, an author, a word, a volume, permitting combination searches, and tagging notations seemed logical choices as core elements. 3) How does one make the data available? Tieing it to a CD-ROM edition creates limitations, but a magnetic tape file that is very large, is accompanied by the encoding specifications, and that allows one to make local modifications also allows one to incorporate any changes one may desire within the bounds of private research, though exporting tag files from a CD-ROM could serve just as well. Since no one on the board could possibly anticipate each and every way in which a scholar might choose to mine this data bank, it was decided to satisfy the basics and make some provisions for what might come. 4) Not to encode the database would rob it of the interchangeability and portability these important texts should accommodate. For CALALUCA, the extensive options presented by full-text searching require care in text selection and strongly support encoding of data to facilitate the widest possible search strategies. Better software can always be created, but summoning the resources, the people, and the energy to reconvert the text is another matter.

PLD is being encoded, captured, and distributed, because to Chadwyck-Healey and the board it offers the widest possible array of future research applications that can be seen today. CALALUCA concluded by urging the encoding of all important text sources in whatever way seems most appropriate and durable at the time, without blanching at the

thought that one's work may require emendation in the future. (Thus, Chadwyck-Healey produced a very large humanities text database before the final release of the TEI Guidelines.)

******

++

DISCUSSION * Creating texts with markup advocated * Trends in encoding * The TEI and the issue of interchangeability of standards * A misconception concerning the TEI * Implications for an institution like LC in the event that a multiplicity of DTDs develops * Producing images as a first step towards possible conversion to full text through character recognition * The AAP tag sets as a common starting point and the need for caution *

++

HOCKEY prefaced the discussion that followed with several comments in favor of creating texts with markup and on trends in encoding. In the future, when many more texts are available for on-line searching, real problems in finding what is wanted will develop, if one is faced with millions of words of data. It therefore becomes important to consider putting markup in texts to help searchers home in on the actual things they wish to retrieve. Various approaches to refining retrieval methods toward this end include building on a computer version of a dictionary and letting the computer look up words in it to obtain more information about the semantic structure or semantic field of a word, its grammatical

structure, and syntactic structure.

HOCKEY commented on the present keen interest in the encoding world in creating: 1) machine-readable versions of dictionaries that can be initially tagged in SGML, which gives a structure to the dictionary entry; these entries can then be converted into a more rigid or otherwise different database structure inside the computer, which can be treated as a dynamic tool for searching mechanisms; 2) large bodies of text to study the language. In order to incorporate more sophisticated mechanisms, more about how words behave needs to be known, which can be learned in part from information in dictionaries. However, the last ten years have seen much interest in studying the structure of printed dictionaries converted into computer-readable form. The information one derives about many words from those is only partial, one or two definitions of the common or the usual meaning of a word, and then numerous definitions of unusual usages. If the computer is using a dictionary to help retrieve words in a text, it needs much more information about the common usages, because those are the ones that occur over and over again. Hence the current interest in developing large bodies of text in computer-readable form in order to study the language. Several projects are engaged in compiling, for example, 100 million words. HOCKEY described one with which she was associated briefly at Oxford University involving compilation of 100 million words of British English: about 10 percent of that will contain detailed linguistic tagging encoded in SGML; it will have word class taggings, with words identified as nouns, verbs, adjectives, or other parts of speech. This tagging can then be used by programs which will begin to learn a bit more about the structure of the

language, and then, can go to tag more text.

HOCKEY said that the more that is tagged accurately, the more one can refine the tagging process and thus the bigger body of text one can build up with linguistic tagging incorporated into it. Hence, the more tagging or annotation there is in the text, the more one may begin to learn about language and the more it will help accomplish more intelligent OCR. She recommended the development of software tools that will help one begin to understand more about a text, which can then be applied to scanning images of that text in that format and to using more intelligence to help one interpret or understand the text.

HOCKEY posited the need to think about common methods of text-encoding for a long time to come, because building these large bodies of text is extremely expensive and will only be done once.

In the more general discussion on approaches to encoding that followed, these points were made:

BESSER identified the underlying problem with standards that all have to struggle with in adopting a standard, namely, the tension between a very highly defined standard that is very interchangeable but does not work for everyone because something is lacking, and a standard that is less defined, more open, more adaptable, but less interchangeable. Contending that the way in which people use SGML is not sufficiently defined, BESSER wondered 1) if people resist the TEI because they think it is too defined

in certain things they do not fit into, and 2) how progress with interchangeability can be made without frightening people away.

SPERBERG-McQUEEN replied that the published drafts of the TEI had met with surprisingly little objection on the grounds that they do not allow one to handle X or Y or Z. Particular concerns of the affiliated projects have led, in practice, to discussions of how extensions are to be made; the primary concern of any project has to be how it can be represented locally, thus making interchange secondary. The TEI has received much criticism based on the notion that everything in it is required or even recommended, which, as it happens, is a misconception from the beginning, because none of it is required and very little is actually actively recommended for all cases, except that one document one's source.

SPERBERG-McQUEEN agreed with BESSER about this trade-off: all the projects in a set of twenty TEI-conformant projects will not necessarily tag the material in the same way. One result of the TEI will be that the easiest problems will be solved--those dealing with the external form of the information; but the problem that is hardest in interchange is that one is not encoding what another wants, and vice versa. Thus, after the adoption of a common notation, the differences in the underlying conceptions of what is interesting about texts become more visible. The success of a standard like the TEI will lie in the ability of the recipient of interchanged texts to use some of what it contains and to add the information that was not encoded that one wants, in a layered way, so that texts can be gradually enriched and one does not

have to put in everything all at once. Hence, having a well-behaved markup scheme is important.

STEVENS followed up on the paradoxical analogy that BESSER alluded to in the example of the MARC records, namely, the formats that are the same except that they are different. STEVENS drew a parallel between document-type definitions and MARC records for books and serials and maps, where one has a tagging structure and there is a text-interchange. STEVENS opined that the producers of the information will set the terms for the standard (i.e., develop document-type definitions for the users of their products), creating a situation that will be problematical for an institution like the Library of Congress, which will have to deal with the DTDs in the event that a multiplicity of them develops. Thus, numerous people are seeking a standard but cannot find the tag set that will be acceptable to them and their clients. SPERBERG-McQUEEN agreed with this view, and said that the situation was in a way worse: attempting to unify arbitrary DTDs resembled attempting to unify a MARC record with a bibliographic record done according to the Prussian instructions. According to STEVENS, this situation occurred very early in the process.

WATERS recalled from early discussions on Project Open Book the concern of many people that merely by producing images, POB was not really enhancing intellectual access to the material. Nevertheless, not wishing to overemphasize the opposition between imaging and full text, WATERS stated that POB views getting the images as a first step toward possibly converting to full text through character recognition, if the technology is appropriate. WATERS also emphasized that encoding is involved even

with a set of images.

SPERBERG-McQUEEN agreed with WATERS that one can create an SGML document consisting wholly of images. At first sight, organizing graphic images with an SGML document may not seem to offer great advantages, but the advantages of the scheme WATERS described would be precisely that ability to move into something that is more of a multimedia document: a combination of transcribed text and page images. WEIBEL concurred in this judgment, offering evidence from Project ADAPT, where a page is divided into text elements and graphic elements, and in fact the text elements are organized by columns and lines. These lines may be used as the basis for distributing documents in a network environment. As one develops software intelligent enough to recognize what those elements are, it makes sense to apply SGML to an image initially, that may, in fact, ultimately become more and more text, either through OCR or edited OCR or even just through keying. For WATERS, the labor of composing the document and saying this set of documents or this set of images belongs to this document constitutes a significant investment.

WEIBEL also made the point that the AAP tag sets, while not excessively prescriptive, offer a common starting point; they do not define the structure of the documents, though. They have some recommendations about DTDs one could use as examples, but they do just suggest tag sets. For example, the CORE project attempts to use the AAP markup as much as possible, but there are clearly areas where structure must be added. That in no way contradicts the use of AAP tag sets.

SPERBERG-McQUEEN noted that the TEI prepared a long working paper early on about the AAP tag set and what it lacked that the TEI thought it needed, and a fairly long critique of the naming conventions, which has led to a very different style of naming in the TEI. He stressed the importance of the opposition between prescriptive markup, the kind that a publisher or anybody can do when producing documents de novo, and descriptive markup, in which one has to take what the text carrier provides. In these particular tag sets it is easy to overemphasize this opposition, because the AAP tag set is extremely flexible. Even if one just used the DTDs, they allow almost anything to appear almost anywhere.

******

SESSION VI. COPYRIGHT ISSUES

++

PETERS * Several cautions concerning copyright in an electronic environment * Review of copyright law in the United States * The notion of the public good and the desirability of incentives to promote it * What copyright protects * Works not protected by copyright * The rights of copyright holders * Publishers' concerns in today's electronic environment * Compulsory licenses * The price of copyright in a digital medium and the need for cooperation * Additional clarifications * Rough justice oftentimes the outcome in numerous copyright matters * Copyright in an electronic society * Copyright law always only sets up the

boundaries; anything can be changed by contract *

++

Marybeth PETERS, policy planning adviser to the Register of Copyrights, Library of Congress, made several general comments and then opened the floor to discussion of subjects of interest to the audience.

Having attended several sessions in an effort to gain a sense of what people did and where copyright would affect their lives, PETERS expressed the following cautions:

* If one takes and converts materials and puts them in new forms, then, from a copyright point of view, one is creating something and will receive some rights.

* However, if what one is converting already exists, a question immediately arises about the status of the materials in question.

* Putting something in the public domain in the United States offers some freedom from anxiety, but distributing it throughout the world on a network is another matter, even if one has put it in the public domain in the United States. Re foreign laws, very frequently a work can be in the public domain in the United States but protected in other countries. Thus, one must consider all of the places a work may reach, lest one unwittingly become liable to being faced

with a suit for copyright infringement, or at least a letter demanding discussion of what one is doing.

PETERS reviewed copyright law in the United States. The U.S. Constitution effectively states that Congress has the power to enact copyright laws for two purposes: 1) to encourage the creation and dissemination of intellectual works for the good of society as a whole; and, significantly, 2) to give creators and those who package and disseminate materials the economic rewards that are due them.

Congress strives to strike a balance, which at times can become an emotional issue. The United States has never accepted the notion of the natural right of an author so much as it has accepted the notion of the public good and the desirability of incentives to promote it. This state of affairs, however, has created strains on the international level and is the reason for several of the differences in the laws that we have. Today the United States protects almost every kind of work that can be called an expression of an author. The standard for gaining copyright protection is simply originality. This is a low standard and means that a work is not copied from something else, as well as shows a certain minimal amount of authorship. One can also acquire copyright protection for making a new version of preexisting material, provided it manifests some spark of creativity.

However, copyright does not protect ideas, methods, systems--only the way that one expresses those things. Nor does copyright protect anything

that is mechanical, anything that does not involve choice, or criteria concerning whether or not one should do a thing. For example, the results of a process called declicking, in which one mechanically removes impure sounds from old recordings, are not copyrightable. On the other hand, the choice to record a song digitally and to increase the sound of violins or to bring up the tympani constitutes the results of conversion that are copyrightable. Moreover, if a work is protected by copyright in the United States, one generally needs the permission of the copyright owner to convert it. Normally, who will own the new--that is, converted--material is a matter of contract. In the absence of a contract, the person who creates the new material is the author and owner. But people do not generally think about the copyright implications until after the fact. PETERS stressed the need when dealing with copyrighted works to think about copyright in advance. One's bargaining power is much greater up front than it is down the road.

PETERS next discussed works not protected by copyright, for example, any work done by a federal employee as part of his or her official duties is in the public domain in the United States. The issue is not wholly free of doubt concerning whether or not the work is in the public domain outside the United States. Other materials in the public domain include: any works published more than seventy-five years ago, and any work published in the United States more than twenty-eight years ago, whose copyright was not renewed. In talking about the new technology and putting material in a digital form to send all over the world, PETERS cautioned, one must keep in mind that while the rights may not be an issue in the United States, they may be in different parts of the world,

where most countries previously employed a copyright term of the life of the author plus fifty years.

PETERS next reviewed the economics of copyright holding. Simply, economic rights are the rights to control the reproduction of a work in any form. They belong to the author, or in the case of a work made for hire, the employer. The second right, which is critical to conversion, is the right to change a work. The right to make new versions is perhaps one of the most significant rights of authors, particularly in an electronic world. The third right is the right to publish the work and the right to disseminate it, something that everyone who deals in an electronic medium needs to know. The basic rule is if a copy is sold, all rights of distribution are extinguished with the sale of that copy. The key is that it must be sold. A number of companies overcome this obstacle by leasing or renting their product. These companies argue that if the material is rented or leased and not sold, they control the uses of a work. The fourth right, and one very important in a digital world, is a right of public performance, which means the right to show the work sequentially. For example, copyright owners control the showing of a CD-ROM product in a public place such as a public library. The reverse side of public performance is something called the right of public display. Moral rights also exist, which at the federal level apply only to very limited visual works of art, but in theory may apply under contract and other principles. Moral rights may include the right of an author to have his or her name on a work, the right of attribution, and the right to object to distortion or mutilation--the right of integrity.

The way copyright law is worded gives much latitude to activities such as preservation; to use of material for scholarly and research purposes when the user does not make multiple copies; and to the generation of facsimile copies of unpublished works by libraries for themselves and other libraries. But the law does not allow anyone to become the distributor of the product for the entire world. In today's electronic environment, publishers are extremely concerned that the entire world is networked and can obtain the information desired from a single copy in a single library. Hence, if there is to be only one sale, which publishers may choose to live with, they will obtain their money in other ways, for example, from access and use. Hence, the development of site licenses and other kinds of agreements to cover what publishers believe they should be compensated for. Any solution that the United States takes today has to consider the international arena.

Noting that the United States is a member of the Berne Convention and subscribes to its provisions, PETERS described the permissions process. She also defined compulsory licenses. A compulsory license, of which the United States has had a few, builds into the law the right to use a work subject to certain terms and conditions. In the international arena, however, the ability to use compulsory licenses is extremely limited. Thus, clearinghouses and other collectives comprise one option that has succeeded in providing for use of a work. Often overlooked when one begins to use copyrighted material and put products together is how expensive the permissions process and managing it is. According to PETERS, the price of copyright in a digital medium, whatever solution is worked out, will include managing and assembling the database. She

strongly recommended that publishers and librarians or people with various backgrounds cooperate to work out administratively feasible systems, in order to produce better results.

In the lengthy question-and-answer period that followed PETERS's presentation, the following points emerged:

* The Copyright Office maintains that anything mechanical and totally exhaustive probably is not protected. In the event that what an individual did in developing potentially copyrightable material is not understood, the Copyright Office will ask about the creative choices the applicant chose to make or not to make. As a practical matter, if one believes she or he has made enough of those choices, that person has a right to assert a copyright and someone else must assert that the work is not copyrightable. The more mechanical, the more automatic, a thing is, the less likely it is to be copyrightable.

* Nearly all photographs are deemed to be copyrightable, but no one worries about them much, because everyone is free to take the same image. Thus, a photographic copyright represents what is called a "thin" copyright. The photograph itself must be duplicated, in order for copyright to be violated.

* The Copyright Office takes the position that X-rays are not copyrightable because they are mechanical. It can be argued

whether or not image enhancement in scanning can be protected. One must exercise care with material created with public funds and generally in the public domain. An article written by a federal employee, if written as part of official duties, is not copyrightable. However, control over a scientific article written by a National Institutes of Health grantee (i.e., someone who receives money from the U.S. government), depends on NIH policy. If the government agency has no policy (and that policy can be contained in its regulations, the contract, or the grant), the author retains copyright. If a provision of the contract, grant, or regulation states that there will be no copyright, then it does not exist. When a work is created, copyright automatically comes into existence unless something exists that says it does not.

* An enhanced electronic copy of a print copy of an older reference work in the public domain that does not contain copyrightable new material is a purely mechanical rendition of the original work, and is not copyrightable.

* Usually, when a work enters the public domain, nothing can remove it. For example, Congress recently passed into law the concept of automatic renewal, which means that copyright on any work published between l964 and l978 does not have to be renewed in order to receive a seventy-five-year term. But any work not renewed before 1964 is in the public domain.

* Concerning whether or not the United States keeps track of when authors die, nothing was ever done, nor is anything being done at the moment by the Copyright Office.

* Software that drives a mechanical process is itself copyrightable. If one changes platforms, the software itself has a copyright. The World Intellectual Property Organization will hold a symposium 28 March through 2 April l993, at Harvard University, on digital technology, and will study this entire issue. If one purchases a computer software package, such as MacPaint, and creates something new, one receives protection only for that which has been added.

PETERS added that often in copyright matters, rough justice is the outcome, for example, in collective licensing, ASCAP (i.e., American Society of Composers, Authors, and Publishers), and BMI (i.e., Broadcast Music, Inc.), where it may seem that the big guys receive more than their due. Of course, people ought not to copy a creative product without paying for it; there should be some compensation. But the truth of the world, and it is not a great truth, is that the big guy gets played on the radio more frequently than the little guy, who has to do much more until he becomes a big guy. That is true of every author, every composer, everyone, and, unfortunately, is part of life.

Copyright always originates with the author, except in cases of works made for hire. (Most software falls into this category.) When an author sends his article to a journal, he has not relinquished copyright, though

he retains the right to relinquish it. The author receives absolutely everything. The less prominent the author, the more leverage the publisher will have in contract negotiations. In order to transfer the rights, the author must sign an agreement giving them away.

In an electronic society, it is important to be able to license a writer and work out deals. With regard to use of a work, it usually is much easier when a publisher holds the rights. In an electronic era, a real problem arises when one is digitizing and making information available. PETERS referred again to electronic licensing clearinghouses. Copyright ought to remain with the author, but as one moves forward globally in the electronic arena, a middleman who can handle the various rights becomes increasingly necessary.

The notion of copyright law is that it resides with the individual, but in an on-line environment, where a work can be adapted and tinkered with by many individuals, there is concern. If changes are authorized and there is no agreement to the contrary, the person who changes a work owns the changes. To put it another way, the person who acquires permission to change a work technically will become the author and the owner, unless some agreement to the contrary has been made. It is typical for the original publisher to try to control all of the versions and all of the uses. Copyright law always only sets up the boundaries. Anything can be changed by contract.

******

SESSION VII. CONCLUSION

++

GENERAL DISCUSSION * Two questions for discussion * Different emphases in the Workshop * Bringing the text and image partisans together * Desiderata in planning the long-term development of something * Questions surrounding the issue of electronic deposit * Discussion of electronic deposit as an allusion to the issue of standards * Need for a directory of preservation projects in digital form and for access to their digitized files * CETH's catalogue of machine-readable texts in the humanities * What constitutes a publication in the electronic world? * Need for LC to deal with the concept of on-line publishing * LC's Network Development Office exploring the limits of MARC as a standard in terms of handling electronic information * Magnitude of the problem and the need for distributed responsibility in order to maintain and store electronic information * Workshop participants to be viewed as a starting point * Development of a network version of AM urged * A step toward AM's construction of some sort of apparatus for network access * A delicate and agonizing policy question for LC * Re the issue of electronic deposit, LC urged to initiate a catalytic process in terms of distributed responsibility * Suggestions for cooperative ventures * Commercial publishers' fears * Strategic questions for getting the image and text people to think through long-term cooperation * Clarification of the driving force behind both the Perseus and the Cornell Xerox projects *

++

In his role as moderator of the concluding session, GIFFORD raised two questions he believed would benefit from discussion: 1) Are there enough commonalities among those of us that have been here for two days so that we can see courses of action that should be taken in the future? And, if so, what are they and who might take them? 2) Partly derivative from that, but obviously very dangerous to LC as host, do you see a role for the Library of Congress in all this? Of course, the Library of Congress holds a rather special status in a number of these matters, because it is not perceived as a player with an economic stake in them, but are there roles that LC can play that can help advance us toward where we are heading?

Describing himself as an uninformed observer of the technicalities of the last two days, GIFFORD detected three different emphases in the Workshop: 1) people who are very deeply committed to text; 2) people who are almost passionate about images; and 3) a few people who are very committed to what happens to the networks. In other words, the new networking dimension, the accessibility of the processability, the portability of all this across the networks. How do we pull those three together?

Adding a question that reflected HOCKEY's comment that this was the fourth workshop she had attended in the previous thirty days, FLEISCHHAUER wondered to what extent this meeting had reinvented the wheel, or if it had contributed anything in the way of bringing together a different group of people from those who normally appear on the workshop circuit.

HOCKEY confessed to being struck at this meeting and the one the Electronic Pierce Consortium organized the previous week that this was a coming together of people working on texts and not images. Attempting to bring the two together is something we ought to be thinking about for the future: How one can think about working with image material to begin with, but structuring it and digitizing it in such a way that at a later stage it can be interpreted into text, and find a common way of building text and images together so that they can be used jointly in the future, with the network support to begin there because that is how people will want to access it.

In planning the long-term development of something, which is what is being done in electronic text, HOCKEY stressed the importance not only of discussing the technical aspects of how one does it but particularly of thinking about what the people who use the stuff will want to do. But conversely, there are numerous things that people start to do with electronic text or material that nobody ever thought of in the beginning.

LESK, in response to the question concerning the role of the Library of Congress, remarked the often suggested desideratum of having electronic deposit: Since everything is now computer-typeset, an entire decade of material that was machine-readable exists, but the publishers frequently did not save it; has LC taken any action to have its copyright deposit operation start collecting these machine-readable versions? In the absence of PETERS, GIFFORD replied that the question was being actively considered but that that was only one dimension of the problem. Another dimension is the whole question of the integrity of the original

electronic document. It becomes highly important in science to prove authorship. How will that be done?

ERWAY explained that, under the old policy, to make a claim for a copyright for works that were published in electronic form, including software, one had to submit a paper copy of the first and last twenty pages of code--something that represented the work but did not include the entire work itself and had little value to anyone. As a temporary measure, LC has claimed the right to demand electronic versions of electronic publications. This measure entails a proactive role for the Library to say that it wants a particular electronic version. Publishers then have perhaps a year to submit it. But the real problem for LC is what to do with all this material in all these different formats. Will the Library mount it? How will it give people access to it? How does LC keep track of the appropriate computers, software, and media? The situation is so hard to control, ERWAY said, that it makes sense for each publishing house to maintain its own archive. But LC cannot enforce that either.

GIFFORD acknowledged LESK's suggestion that establishing a priority offered the solution, albeit a fairly complicated one. But who maintains that register?, he asked. GRABER noted that LC does attempt to collect a Macintosh version and the IBM-compatible version of software. It does not collect other versions. But while true for software, BYRUM observed, this reply does not speak to materials, that is, all the materials that were published that were on somebody's microcomputer or driver tapes at a publishing office across the country. LC does well to acquire specific machine-readable products selectively that were intended to be

machine-readable. Materials that were in machine-readable form at one time, BYRUM said, would be beyond LC's capability at the moment, insofar as attempting to acquire, organize, and preserve them are concerned--and preservation would be the most important consideration. In this connection, GIFFORD reiterated the need to work out some sense of distributive responsibility for a number of these issues, which inevitably will require significant cooperation and discussion. Nobody can do it all.

LESK suggested that some publishers may look with favor on LC beginning to serve as a depository of tapes in an electronic manuscript standard. Publishers may view this as a service that they did not have to perform and they might send in tapes. However, SPERBERG-McQUEEN countered, although publishers have had equivalent services available to them for a long time, the electronic text archive has never turned away or been flooded with tapes and is forever sending feedback to the depositor. Some publishers do send in tapes.

ANDRE viewed this discussion as an allusion to the issue of standards. She recommended that the AAP standard and the TEI, which has already been somewhat harmonized internationally and which also shares several compatibilities with the AAP, be harmonized to ensure sufficient compatibility in the software. She drew the line at saying LC ought to be the locus or forum for such harmonization.

Taking the group in a slightly different direction, but one where at

least in the near term LC might play a helpful role, LYNCH remarked the plans of a number of projects to carry out preservation by creating digital images that will end up in on-line or near-line storage at some institution. Presumably, LC will link this material somehow to its on-line catalog in most cases. Thus, it is in a digital form. LYNCH had the impression that many of these institutions would be willing to make those files accessible to other people outside the institution, provided that there is no copyright problem. This desideratum will require propagating the knowledge that those digitized files exist, so that they can end up in other on-line catalogs. Although uncertain about the mechanism for achieving this result, LYNCH said that it warranted scrutiny because it seemed to be connected to some of the basic issues of cataloging and distribution of records. It would be foolish, given the amount of work that all of us have to do and our meager resources, to discover multiple institutions digitizing the same work. Re microforms, LYNCH said, we are in pretty good shape.

BATTIN called this a big problem and noted that the Cornell people (who had already departed) were working on it. At issue from the beginning was to learn how to catalog that information into RLIN and then into OCLC, so that it would be accessible. That issue remains to be resolved. LYNCH rejoined that putting it into OCLC or RLIN was helpful insofar as somebody who is thinking of performing preservation activity on that work could learn about it. It is not necessarily helpful for institutions to make that available. BATTIN opined that the idea was that it not only be for preservation purposes but for the convenience of people looking for this material. She endorsed LYNCH's dictum that duplication of this

effort was to be avoided by every means.

HOCKEY informed the Workshop about one major current activity of CETH, namely a catalogue of machine-readable texts in the humanities. Held on RLIN at present, the catalogue has been concentrated on ASCII as opposed to digitized images of text. She is exploring ways to improve the catalogue and make it more widely available, and welcomed suggestions about these concerns. CETH owns the records, which are not just restricted to RLIN, and can distribute them however it wishes.

Taking up LESK's earlier question, BATTIN inquired whether LC, since it is accepting electronic files and designing a mechanism for dealing with that rather than putting books on shelves, would become responsible for the National Copyright Depository of Electronic Materials. Of course that could not be accomplished overnight, but it would be something LC could plan for. GIFFORD acknowledged that much thought was being devoted to that set of problems and returned the discussion to the issue raised by LYNCH--whether or not putting the kind of records that both BATTIN and HOCKEY have been talking about in RLIN is not a satisfactory solution. It seemed to him that RLIN answered LYNCH's original point concerning some kind of directory for these kinds of materials. In a situation where somebody is attempting to decide whether or not to scan this or film that or to learn whether or not someone has already done so, LYNCH suggested, RLIN is helpful, but it is not helpful in the case of a local, on-line catalogue. Further, one would like to have her or his system be aware that that exists in digital form, so that one can present it to a patron, even though one did not digitize it, if it is out of copyright.

The only way to make those linkages would be to perform a tremendous amount of real-time look-up, which would be awkward at best, or periodically to yank the whole file from RLIN and match it against one's own stuff, which is a nuisance.

But where, ERWAY inquired, does one stop including things that are available with Internet, for instance, in one's local catalogue? It almost seems that that is LC's means to acquire access to them. That represents LC's new form of library loan. Perhaps LC's new on-line catalogue is an amalgamation of all these catalogues on line. LYNCH conceded that perhaps that was true in the very long term, but was not applicable to scanning in the short term. In his view, the totals cited by Yale, 10,000 books over perhaps a four-year period, and 1,000-1,500 books from Cornell, were not big numbers, while searching all over creation for relatively rare occurrences will prove to be less efficient. As GIFFORD wondered if this would not be a separable file on RLIN and could be requested from them, BATTIN interjected that it was easily accessible to an institution. SEVERTSON pointed out that that file, cum enhancements, was available with reference information on CD-ROM, which makes it a little more available.

In HOCKEY's view, the real question facing the Workshop is what to put in this catalogue, because that raises the question of what constitutes a publication in the electronic world. (WEIBEL interjected that Eric Joule in OCLC's Office of Research is also wrestling with this particular problem, while GIFFORD thought it sounded fairly generic.) HOCKEY contended that a majority of texts in the humanities are in the hands

of either a small number of large research institutions or individuals and are not generally available for anyone else to access at all. She wondered if these texts ought to be catalogued.

After argument proceeded back and forth for several minutes over why cataloguing might be a necessary service, LEBRON suggested that this issue involved the responsibility of a publisher. The fact that someone has created something electronically and keeps it under his or her control does not constitute publication. Publication implies dissemination. While it would be important for a scholar to let other people know that this creation exists, in many respects this is no different from an unpublished manuscript. That is what is being accessed in there, except that now one is not looking at it in the hard-copy but in the electronic environment.

LEBRON expressed puzzlement at the variety of ways electronic publishing has been viewed. Much of what has been discussed throughout these two days has concerned CD-ROM publishing, whereas in the on-line environment that she confronts, the constraints and challenges are very different. Sooner or later LC will have to deal with the concept of on-line publishing. Taking up the comment ERWAY made earlier about storing copies, LEBRON gave her own journal as an example. How would she deposit OJCCT for copyright?, she asked, because the journal will exist in the mainframe at OCLC and people will be able to access it. Here the situation is different, ownership versus access, and is something that arises with publication in the on-line environment, faster than is sometimes realized. Lacking clear answers to all of these questions

herself, LEBRON did not anticipate that LC would be able to take a role in helping to define some of them for quite a while.

GREENFIELD observed that LC's Network Development Office is attempting, among other things, to explore the limits of MARC as a standard in terms of handling electronic information. GREENFIELD also noted that Rebecca GUENTHER from that office gave a paper to the American Society for Information Science (ASIS) summarizing several of the discussion papers that were coming out of the Network Development Office. GREENFIELD said he understood that that office had a list-server soliciting just the kind of feedback received today concerning the difficulties of identifying and cataloguing electronic information. GREENFIELD hoped that everybody would be aware of that and somehow contribute to that conversation.

Noting two of LC's roles, first, to act as a repository of record for material that is copyrighted in this country, and second, to make materials it holds available in some limited form to a clientele that goes beyond Congress, BESSER suggested that it was incumbent on LC to extend those responsibilities to all the things being published in electronic form. This would mean eventually accepting electronic formats. LC could require that at some point they be in a certain limited set of formats, and then develop mechanisms for allowing people to access those in the same way that other things are accessed. This does not imply that they are on the network and available to everyone. LC does that with most of its bibliographic records, BESSER said, which end up migrating to the utility (e.g., OCLC) or somewhere else. But just as most of LC's books are available in some form through interlibrary

loan or some other mechanism, so in the same way electronic formats ought to be available to others in some format, though with some copyright considerations. BESSER was not suggesting that these mechanisms be established tomorrow, only that they seemed to fall within LC's purview, and that there should be long-range plans to establish them.

Acknowledging that those from LC in the room agreed with BESSER concerning the need to confront difficult questions, GIFFORD underscored the magnitude of the problem of what to keep and what to select. GIFFORD noted that LC currently receives some 31,000 items per day, not counting electronic materials, and argued for much more distributed responsibility in order to maintain and store electronic information.

BESSER responded that the assembled group could be viewed as a starting point, whose initial operating premise could be helping to move in this direction and defining how LC could do so, for example, in areas of standardization or distribution of responsibility.

FLEISCHHAUER added that AM was fully engaged, wrestling with some of the questions that pertain to the conversion of older historical materials, which would be one thing that the Library of Congress might do. Several points mentioned by BESSER and several others on this question have a much greater impact on those who are concerned with cataloguing and the networking of bibliographic information, as well as preservation itself.

Speaking directly to AM, which he considered was a largely uncopyrighted

database, LYNCH urged development of a network version of AM, or consideration of making the data in it available to people interested in doing network multimedia. On account of the current great shortage of digital data that is both appealing and unencumbered by complex rights problems, this course of action could have a significant effect on making network multimedia a reality.

In this connection, FLEISCHHAUER reported on a fragmentary prototype in LC's Office of Information Technology Services that attempts to associate digital images of photographs with cataloguing information in ways that work within a local area network--a step, so to say, toward AM's construction of some sort of apparatus for access. Further, AM has attempted to use standard data forms in order to help make that distinction between the access tools and the underlying data, and thus believes that the database is networkable.

A delicate and agonizing policy question for LC, however, which comes back to resources and unfortunately has an impact on this, is to find some appropriate, honorable, and legal cost-recovery possibilities. A certain skittishness concerning cost-recovery has made people unsure exactly what to do. AM would be highly receptive to discussing further LYNCH's offer to test or demonstrate its database in a network environment, FLEISCHHAUER said.

Returning the discussion to what she viewed as the vital issue of electronic deposit, BATTIN recommended that LC initiate a catalytic

process in terms of distributed responsibility, that is, bring together the distributed organizations and set up a study group to look at all these issues and see where we as a nation should move. The broader issues of how we deal with the management of electronic information will not disappear, but only grow worse.

LESK took up this theme and suggested that LC attempt to persuade one major library in each state to deal with its state equivalent publisher, which might produce a cooperative project that would be equitably distributed around the country, and one in which LC would be dealing with a minimal number of publishers and minimal copyright problems.

GRABER remarked the recent development in the scientific community of a willingness to use SGML and either deposit or interchange on a fairly standardized format. He wondered if a similar movement was taking place in the humanities. Although the National Library of Medicine found only a few publishers to cooperate in a like venture two or three years ago, a new effort might generate a much larger number willing to cooperate.

KIMBALL recounted his unit's (Machine-Readable Collections Reading Room) troubles with the commercial publishers of electronic media in acquiring materials for LC's collections, in particular the publishers' fear that they would not be able to cover their costs and would lose control of their products, that LC would give them away or sell them and make profits from them. He doubted that the publishing industry was prepared to move into this area at the moment, given its resistance to allowing LC

to use its machine-readable materials as the Library would like.

The copyright law now addresses compact disk as a medium, and LC can request one copy of that, or two copies if it is the only version, and can request copies of software, but that fails to address magazines or books or anything like that which is in machine-readable form.

GIFFORD acknowledged the thorny nature of this issue, which he illustrated with the example of the cumbersome process involved in putting a copy of a scientific database on a LAN in LC's science reading room. He also acknowledged that LC needs help and could enlist the energies and talents of Workshop participants in thinking through a number of these problems.

GIFFORD returned the discussion to getting the image and text people to think through together where they want to go in the long term. MYLONAS conceded that her experience at the Pierce Symposium the previous week at Georgetown University and this week at LC had forced her to reevaluate her perspective on the usefulness of text as images. MYLONAS framed the issues in a series of questions: How do we acquire machine-readable text? Do we take pictures of it and perform OCR on it later? Is it important to obtain very high-quality images and text, etc.?
FLEISCHHAUER agreed with MYLONAS's framing of strategic questions, adding that a large institution such as LC probably has to do all of those things at different times. Thus, the trick is to exercise judgment. The Workshop had added to his and AM's considerations in making those judgments. Concerning future meetings or discussions, MYLONAS suggested

that screening priorities would be helpful.

WEIBEL opined that the diversity reflected in this group was a sign both of the health and of the immaturity of the field, and more time would have to pass before we convince one another concerning standards.

An exchange between MYLONAS and BATTIN clarified the point that the driving force behind both the Perseus and the Cornell Xerox projects was the preservation of knowledge for the future, not simply for particular research use. In the case of Perseus, MYLONAS said, the assumption was that the texts would not be entered again into electronically readable form. SPERBERG-McQUEEN added that a scanned image would not serve as an archival copy for purposes of preservation in the case of, say, the Bill of Rights, in the sense that the scanned images are effectively the archival copies for the Cornell mathematics books.

*** *** *** ****** *** *** ***

Appendix I: PROGRAM

WORKSHOP
ON
ELECTRONIC
TEXTS

9-10 June 1992

Library of Congress

Washington, D.C.

Supported by a Grant from the David and Lucile Packard Foundation

Tuesday, 9 June 1992

NATIONAL DEMONSTRATION LAB, ATRIUM, LIBRARY MADISON

8:30 AM Coffee and Danish, registration

9:00 AM Welcome

Prosser Gifford, Director for Scholarly Programs, and Carl

Fleischhauer, Coordinator, American Memory, Library of

Congress

9:l5 AM Session I. Content in a New Form: Who Will Use It and What

Will They Do?

Broad description of the range of electronic information.

Characterization of who uses it and how it is or may be used.
In addition to a look at scholarly uses, this session will
include a presentation on use by students (K-12 and college)
and the general public.

Moderator: James Daly
Avra Michelson, Archival Research and Evaluation Staff,
National Archives and Records Administration (Overview)
Susan H. Veccia, Team Leader, American Memory, User Evaluation,
and
Joanne Freeman, Associate Coordinator, American Memory, Library
of Congress (Beyond the scholar)

10:30-
11:00 AM Break

11:00 AM Session II. Show and Tell.

Each presentation to consist of a fifteen-minute
statement/show; group discussion will follow lunch.

Moderator: Jacqueline Hess, Director, National Demonstration
Lab

1. A classics project, stressing texts and text retrieval

more than multimedia: Perseus Project, Harvard

University

Elli Mylonas, Managing Editor

2. Other humanities projects employing the emerging norms of

the Text Encoding Initiative (TEI): Chadwyck-Healey's

The English Poetry Full Text Database and/or Patrologia

Latina Database

Eric M. Calaluca, Vice President, Chadwyck-Healey, Inc.

3. American Memory

Carl Fleischhauer, Coordinator, and

Ricky Erway, Associate Coordinator, Library of Congress

4. Founding Fathers example from Packard Humanities

Institute: The Papers of George Washington, University

of Virginia

Dorothy Twohig, Managing Editor, and/or

David Woodley Packard

5. An electronic medical journal offering graphics and

full-text searchability: The Online Journal of Current

Clinical Trials, American Association for the Advancement

of Science

Maria L. Lebron, Managing Editor

6. A project that offers facsimile images of pages but omits

searchable text: Cornell math books

Lynne K. Personius, Assistant Director, Cornell

Information Technologies for Scholarly Information

Sources, Cornell University

12:30 PM Lunch (Dining Room A, Library Madison 620. Exhibits

available.)

1:30 PM Session II. Show and Tell (Cont'd.).

3:00-

3:30 PM Break

3:30-

5:30 PM Session III. Distribution, Networks, and Networking: Options

for Dissemination.

Published disks: University presses and public-sector

publishers, private-sector publishers

Computer networks

Moderator: Robert G. Zich, Special Assistant to the Associate

Librarian for Special Projects, Library of Congress

Clifford A. Lynch, Director, Library Automation, University of California

Howard Besser, School of Library and Information Science, University of Pittsburgh

Ronald L. Larsen, Associate Director of Libraries for Information Technology, University of Maryland at College Park

Edwin B. Brownrigg, Executive Director, Memex Research Institute

6:30 PM Reception (Montpelier Room, Library Madison 619.)

******

Wednesday, 10 June 1992

DINING ROOM A, LIBRARY MADISON 620

8:30 AM Coffee and Danish

9:00 AM Session IV. Image Capture, Text Capture, Overview of Text and Image Storage Formats.

Moderator: William L. Hooton, Vice President of Operations, I-NET

A) Principal Methods for Image Capture of Text:

Direct scanning

Use of microform

Anne R. Kenney, Assistant Director, Department of Preservation and Conservation, Cornell University

Pamela Q.J. Andre, Associate Director, Automation, and Judith A. Zidar, Coordinator, National Agricultural Text Digitizing Program (NATDP), National Agricultural Library (NAL)

Donald J. Waters, Head, Systems Office, Yale University Library

B) Special Problems:

Bound volumes

Conservation

Reproducing printed halftones

Carl Fleischhauer, Coordinator, American Memory, Library of Congress

George Thoma, Chief, Communications Engineering Branch, National Library of Medicine (NLM)

10:30-

11:00 AM Break

11:00 AM Session IV. Image Capture, Text Capture, Overview of Text and Image Storage Formats (Cont'd.).

C) Image Standards and Implications for Preservation

Jean Baronas, Senior Manager, Department of Standards and Technology, Association for Information and Image Management (AIIM)

Patricia Battin, President, The Commission on Preservation and Access (CPA)

D) Text Conversion:

OCR vs. rekeying

Standards of accuracy and use of imperfect texts

Service bureaus

Stuart Weibel, Senior Research Specialist, Online Computer Library Center, Inc. (OCLC)

Michael Lesk, Executive Director, Computer Science Research, Bellcore

Ricky Erway, Associate Coordinator, American Memory, Library of Congress

Pamela Q.J. Andre, Associate Director, Automation, and Judith A. Zidar, Coordinator, National Agricultural Text Digitizing Program (NATDP), National Agricultural Library

(NAL)

12:30-

1:30 PM Lunch

1:30 PM Session V. Approaches to Preparing Electronic Texts.

Discussion of approaches to structuring text for the computer;

pros and cons of text coding, description of methods in

practice, and comparison of text-coding methods.

Moderator: Susan Hockey, Director, Center for Electronic Texts

in the Humanities (CETH), Rutgers and Princeton Universities

David Woodley Packard

C.M. Sperberg-McQueen, Editor, Text Encoding Initiative (TEI),

University of Illinois-Chicago

Eric M. Calaluca, Vice President, Chadwyck-Healey, Inc.

3:30-

4:00 PM Break

4:00 PM Session VI. Copyright Issues.

Marybeth Peters, Policy Planning Adviser to the Register of

Copyrights, Library of Congress

5:00 PM Session VII. Conclusion.

General discussion.

What topics were omitted or given short shrift that anyone would like to talk about now?

Is there a "group" here? What should the group do next, if anything? What should the Library of Congress do next, if anything?

Moderator: Prosser Gifford, Director for Scholarly Programs, Library of Congress

6:00 PM Adjourn

*** *** *** ****** *** *** ***

Appendix II: ABSTRACTS

SESSION I

Avra MICHELSON Forecasting the Use of Electronic Texts by Social Sciences and Humanities Scholars

This presentation explores the ways in which electronic texts are likely to be used by the non-scientific scholarly community. Many of the remarks are drawn from a report the speaker coauthored with Jeff Rothenberg, a computer scientist at The RAND Corporation.

The speaker assesses 1) current scholarly use of information technology and 2) the key trends in information technology most relevant to the research process, in order to predict how social sciences and humanities scholars are apt to use electronic texts. In introducing the topic, current use of electronic texts is explored broadly within the context of scholarly communication. From the perspective of scholarly communication, the work of humanities and social sciences scholars involves five processes: 1) identification of sources, 2) communication with colleagues, 3) interpretation and analysis of data, 4) dissemination of research findings, and 5) curriculum development and instruction. The extent to which computation currently permeates aspects of scholarly communication represents a viable indicator of the prospects for electronic texts.

The discussion of current practice is balanced by an analysis of key trends in the scholarly use of information technology. These include the trends toward end-user computing and connectivity, which provide a framework for forecasting the use of electronic texts through this millennium. The presentation concludes with a summary of the ways in which the nonscientific scholarly community can be expected to use electronic texts, and the implications of that use for information

providers.

Susan VECCIA and Joanne FREEMAN Electronic Archives for the Public: Use of American Memory in Public and School Libraries

This joint discussion focuses on nonscholarly applications of electronic library materials, specifically addressing use of the Library of Congress American Memory (AM) program in a small number of public and school libraries throughout the United States. AM consists of selected Library of Congress primary archival materials, stored on optical media (CD-ROM/videodisc), and presented with little or no editing. Many collections are accompanied by electronic introductions and user's guides offering background information and historical context. Collections represent a variety of formats including photographs, graphic arts, motion pictures, recorded sound, music, broadsides and manuscripts, books, and pamphlets.

In 1991, the Library of Congress began a nationwide evaluation of AM in different types of institutions. Test sites include public libraries, elementary and secondary school libraries, college and university libraries, state libraries, and special libraries. Susan VECCIA and Joanne FREEMAN will discuss their observations on the use of AM by the nonscholarly community, using evidence gleaned from this ongoing evaluation effort.

VECCIA will comment on the overall goals of the evaluation project, and the types of public and school libraries included in this study. Her comments on nonscholarly use of AM will focus on the public library as a cultural and community institution, often bridging the gap between formal and informal education. FREEMAN will discuss the use of AM in school libraries. Use by students and teachers has revealed some broad questions about the use of electronic resources, as well as definite benefits gained by the "nonscholar." Topics will include the problem of grasping content and context in an electronic environment, the stumbling blocks created by "new" technologies, and the unique skills and interests awakened through use of electronic resources.

SESSION II

Elli MYLONAS The Perseus Project: Interactive Sources and Studies in Classical Greece

The Perseus Project (5) has just released Perseus 1.0, the first publicly available version of its hypertextual database of multimedia materials on classical Greece. Perseus is designed to be used by a wide audience, comprised of readers at the student and scholar levels. As such, it must be able to locate information using different strategies, and it must contain enough detail to serve the different needs of its users. In addition, it must be delivered so that it is affordable to its target audience. [These problems and the solutions we chose are described in Mylonas, "An Interface to Classical Greek Civilization," JASIS 43:2,

March 1992.]

In order to achieve its objective, the project staff decided to make a conscious separation between selecting and converting textual, database, and image data on the one hand, and putting it into a delivery system on the other. That way, it is possible to create the electronic data without thinking about the restrictions of the delivery system. We have made a great effort to choose system-independent formats for our data, and to put as much thought and work as possible into structuring it so that the translation from paper to electronic form will enhance the value of the data. [A discussion of these solutions as of two years ago is in Elli Mylonas, Gregory Crane, Kenneth Morrell, and D. Neel Smith, "The Perseus Project: Data in the Electronic Age," in Accessing Antiquity: The Computerization of Classical Databases, J. Solomon and T. Worthen (eds.), University of Arizona Press, in press.]

Much of the work on Perseus is focused on collecting and converting the data on which the project is based. At the same time, it is necessary to provide means of access to the information, in order to make it usable, and them to investigate how it is used. As we learn more about what students and scholars from different backgrounds do with Perseus, we can adjust our data collection, and also modify the system to accommodate them. In creating a delivery system for general use, we have tried to avoid favoring any one type of use by allowing multiple forms of access to and navigation through the system.

The way text is handled exemplifies some of these principles. All text in Perseus is tagged using SGML, following the guidelines of the Text Encoding Initiative (TEI). This markup is used to index the text, and process it so that it can be imported into HyperCard. No SGML markup remains in the text that reaches the user, because currently it would be too expensive to create a system that acts on SGML in real time. However, the regularity provided by SGML is essential for verifying the content of the texts, and greatly speeds all the processing performed on them. The fact that the texts exist in SGML ensures that they will be relatively easy to port to different hardware and software, and so will outlast the current delivery platform. Finally, the SGML markup incorporates existing canonical reference systems (chapter, verse, line, etc.); indexing and navigation are based on these features. This ensures that the same canonical reference will always resolve to the same point within a text, and that all versions of our texts, regardless of delivery platform (even paper printouts) will function the same way.

In order to provide tools for users, the text is processed by a morphological analyzer, and the results are stored in a database. Together with the index, the Greek-English Lexicon, and the index of all the English words in the definitions of the lexicon, the morphological analyses comprise a set of linguistic tools that allow users of all levels to work with the textual information, and to accomplish different tasks. For example, students who read no Greek may explore a concept as it appears in Greek texts by using the English-Greek index, and then looking up works in the texts and translations, or scholars may do detailed morphological studies of word use by using the morphological

analyses of the texts. Because these tools were not designed for any one use, the same tools and the same data can be used by both students and scholars.

NOTES:

(5) Perseus is based at Harvard University, with collaborators at several other universities. The project has been funded primarily by the Annenberg/CPB Project, as well as by Harvard University, Apple Computer, and others. It is published by Yale University Press. Perseus runs on Macintosh computers, under the HyperCard program.

Eric CALALUCA

Chadwyck-Healey embarked last year on two distinct yet related full-text humanities database projects.

The English Poetry Full-Text Database and the Patrologia Latina Database represent new approaches to linguistic research resources. The size and complexity of the projects present problems for electronic publishers, but surmountable ones if they remain abreast of the latest possibilities in data capture and retrieval software techniques.

The issues which required address prior to the commencement of the projects were legion:

1. Editorial selection (or exclusion) of materials in each database

2. Deciding whether or not to incorporate a normative encoding structure into the databases?

A. If one is selected, should it be SGML?

B. If SGML, then the TEI?

3. Deliver as CD-ROM, magnetic tape, or both?

4. Can one produce retrieval software advanced enough for the postdoctoral linguist, yet accessible enough for unattended general use? Should one try?

5. Re fair and liberal networking policies, what are the risks to an electronic publisher?

6. How does the emergence of national and international education networks affect the use and viability of research projects requiring high investment? Do the new European Community directives concerning database protection necessitate two distinct publishing projects, one for North America and one for overseas?

From new notions of "scholarly fair use" to the future of optical media, virtually every issue related to electronic publishing was aired. The result is two projects which have been constructed to provide the quality research resources with the fewest encumbrances to use by teachers and private scholars.

Dorothy TWOHIG

In spring 1988 the editors of the papers of George Washington, John Adams, Thomas Jefferson, James Madison, and Benjamin Franklin were approached by classics scholar David Packard on behalf of the Packard Humanities Foundation with a proposal to produce a CD-ROM edition of the complete papers of each of the Founding Fathers. This electronic edition will supplement the published volumes, making the documents widely available to students and researchers at reasonable cost. We estimate that our CD-ROM edition of Washington's Papers will be substantially completed within the next two years and ready for publication. Within the next ten years or so, similar CD-ROM editions of the Franklin, Adams, Jefferson, and Madison papers also will be available. At the Library of Congress's session on technology, I would like to discuss not only the experience of the Washington Papers in producing the CD-ROM edition, but the impact technology has had on these major editorial projects. Already, we are editing our volumes with an eye to the material that will be readily available in the CD-ROM edition. The completed electronic edition will provide immense possibilities for the searching of documents for information in a way never possible before. The kind of technical innovations that are currently available and on the drawing board will

soon revolutionize historical research and the production of historical documents. Unfortunately, much of this new technology is not being used in the planning stages of historical projects, simply because many historians are aware only in the vaguest way of its existence. At least two major new historical editing projects are considering microfilm editions, simply because they are not aware of the possibilities of electronic alternatives and the advantages of the new technology in terms of flexibility and research potential compared to microfilm. In fact, too many of us in history and literature are still at the stage of struggling with our PCs. There are many historical editorial projects in progress presently, and an equal number of literary projects. While the two fields have somewhat different approaches to textual editing, there are ways in which electronic technology can be of service to both.

Since few of the editors involved in the Founding Fathers CD-ROM editions are technical experts in any sense, I hope to point out in my discussion of our experience how many of these electronic innovations can be used successfully by scholars who are novices in the world of new technology. One of the major concerns of the sponsors of the multitude of new scholarly editions is the limited audience reached by the published volumes. Most of these editions are being published in small quantities and the publishers' price for them puts them out of the reach not only of individual scholars but of most public libraries and all but the largest educational institutions. However, little attention is being given to ways in which technology can bypass conventional publication to make historical and literary documents more widely available.

What attracted us most to the CD-ROM edition of The Papers of George Washington was the fact that David Packard's aim was to make a complete edition of all of the 135,000 documents we have collected available in an inexpensive format that would be placed in public libraries, small colleges, and even high schools. This would provide an audience far beyond our present 1,000-copy, $45 published edition. Since the CD-ROM edition will carry none of the explanatory annotation that appears in the published volumes, we also feel that the use of the CD-ROM will lead many researchers to seek out the published volumes.

In addition to ignorance of new technical advances, I have found that too many editors--and historians and literary scholars--are resistant and even hostile to suggestions that electronic technology may enhance their work. I intend to discuss some of the arguments traditionalists are advancing to resist technology, ranging from distrust of the speed with which it changes (we are already wondering what is out there that is better than CD-ROM) to suspicion of the technical language used to describe electronic developments.

Maria LEBRON

The Online Journal of Current Clinical Trials, a joint venture of the American Association for the Advancement of Science (AAAS) and the Online Computer Library Center, Inc. (OCLC), is the first peer-reviewed journal to provide full text, tabular material, and line illustrations on line.
This presentation will discuss the genesis and start-up period of the

journal. Topics of discussion will include historical overview, day-to-day management of the editorial peer review, and manuscript tagging and publication. A demonstration of the journal and its features will accompany the presentation.

Lynne PERSONIUS

Cornell University Library, Cornell Information Technologies, and Xerox Corporation, with the support of the Commission on Preservation and Access, and Sun Microsystems, Inc., have been collaborating in a project to test a prototype system for recording brittle books as digital images and producing, on demand, high-quality archival paper replacements. The project goes beyond that, however, to investigate some of the issues surrounding scanning, storing, retrieving, and providing access to digital images in a network environment.

The Joint Study in Digital Preservation began in January 1990. Xerox provided the College Library Access and Storage System (CLASS) software, a prototype 600-dots-per-inch (dpi) scanner, and the hardware necessary to support network printing on the DocuTech printer housed in Cornell's Computing and Communications Center (CCC).

The Cornell staff using the hardware and software became an integral part of the development and testing process for enhancements to the CLASS software system. The collaborative nature of this relationship is resulting in a system that is specifically tailored to the preservation

application.

A digital library of 1,000 volumes (or approximately 300,000 images) has been created and is stored on an optical jukebox that resides in CCC. The library includes a collection of select mathematics monographs that provides mathematics faculty with an opportunity to use the electronic library. The remaining volumes were chosen for the library to test the various capabilities of the scanning system.

One project objective is to provide users of the Cornell library and the library staff with the ability to request facsimiles of digitized images or to retrieve the actual electronic image for browsing. A prototype viewing workstation has been created by Xerox, with input into the design by a committee of Cornell librarians and computer professionals. This will allow us to experiment with patron access to the images that make up the digital library. The viewing station provides search, retrieval, and (ultimately) printing functions with enhancements to facilitate navigation through multiple documents.

Cornell currently is working to extend access to the digital library to readers using workstations from their offices. This year is devoted to the development of a network resident image conversion and delivery server, and client software that will support readers who use Apple Macintosh computers, IBM windows platforms, and Sun workstations. Equipment for this development was provided by Sun Microsystems with support from the Commission on Preservation and Access.

During the show-and-tell session of the Workshop on Electronic Texts, a prototype view station will be demonstrated. In addition, a display of original library books that have been digitized will be available for review with associated printed copies for comparison. The fifteen-minute overview of the project will include a slide presentation that constitutes a "tour" of the preservation digitizing process.

The final network-connected version of the viewing station will provide library users with another mechanism for accessing the digital library, and will also provide the capability of viewing images directly. This will not require special software, although a powerful computer with good graphics will be needed.

The Joint Study in Digital Preservation has generated a great deal of interest in the library community. Unfortunately, or perhaps fortunately, this project serves to raise a vast number of other issues surrounding the use of digital technology for the preservation and use of deteriorating library materials, which subsequent projects will need to examine. Much work remains.

SESSION III

Howard BESSER Networking Multimedia Databases

What do we have to consider in building and distributing databases of visual materials in a multi-user environment? This presentation examines a variety of concerns that need to be addressed before a multimedia database can be set up in a networked environment.

In the past it has not been feasible to implement databases of visual materials in shared-user environments because of technological barriers. Each of the two basic models for multi-user multimedia databases has posed its own problem. The analog multimedia storage model (represented by Project Athena's parallel analog and digital networks) has required an incredibly complex (and expensive) infrastructure. The economies of scale that make multi-user setups cheaper per user served do not operate in an environment that requires a computer workstation, videodisc player, and two display devices for each user.

The digital multimedia storage model has required vast amounts of storage space (as much as one gigabyte per thirty still images). In the past the cost of such a large amount of storage space made this model a prohibitive choice as well. But plunging storage costs are finally making this second alternative viable.

If storage no longer poses such an impediment, what do we need to consider in building digitally stored multi-user databases of visual materials? This presentation will examine the networking and telecommunication constraints that must be overcome before such databases can become commonplace and useful to a large number of people.

The key problem is the vast size of multimedia documents, and how this affects not only storage but telecommunications transmission time. Anything slower than T-1 speed is impractical for files of 1 megabyte or larger (which is likely to be small for a multimedia document). For instance, even on a 56 Kb line it would take three minutes to transfer a 1-megabyte file. And these figures assume ideal circumstances, and do not take into consideration other users contending for network bandwidth, disk access time, or the time needed for remote display. Current common telephone transmission rates would be completely impractical; few users would be willing to wait the hour necessary to transmit a single image at 2400 baud.

This necessitates compression, which itself raises a number of other issues. In order to decrease file sizes significantly, we must employ lossy compression algorithms. But how much quality can we afford to lose? To date there has been only one significant study done of image-quality needs for a particular user group, and this study did not look at loss resulting from compression. Only after identifying image-quality needs can we begin to address storage and network bandwidth needs.

Experience with X-Windows-based applications (such as Imagequery, the University of California at Berkeley image database) demonstrates the utility of a client-server topology, but also points to the limitation of current software for a distributed environment. For example,

applications like Imagequery can incorporate compression, but current X implementations do not permit decompression at the end user's workstation. Such decompression at the host computer alleviates storage capacity problems while doing nothing to address problems of telecommunications bandwidth.

We need to examine the effects on network through-put of moving multimedia documents around on a network. We need to examine various topologies that will help us avoid bottlenecks around servers and gateways. Experience with applications such as these raise still broader questions. How closely is the multimedia document tied to the software for viewing it? Can it be accessed and viewed from other applications? Experience with the MARC format (and more recently with the Z39.50 protocols) shows how useful it can be to store documents in a form in which they can be accessed by a variety of application software.

Finally, from an intellectual-access standpoint, we need to address the issue of providing access to these multimedia documents in interdisciplinary environments. We need to examine terminology and indexing strategies that will allow us to provide access to this material in a cross-disciplinary way.

Ronald LARSEN Directions in High-Performance Networking for Libraries

The pace at which computing technology has advanced over the past forty

years shows no sign of abating. Roughly speaking, each five-year period has yielded an order-of-magnitude improvement in price and performance of computing equipment. No fundamental hurdles are likely to prevent this pace from continuing for at least the next decade. It is only in the past five years, though, that computing has become ubiquitous in libraries, affecting all staff and patrons, directly or indirectly.

During these same five years, communications rates on the Internet, the principal academic computing network, have grown from 56 kbps to 1.5 Mbps, and the NSFNet backbone is now running 45 Mbps. Over the next five years, communication rates on the backbone are expected to exceed 1 Gbps. Growth in both the population of network users and the volume of network traffic has continued to grow geometrically, at rates approaching 15 percent per month. This flood of capacity and use, likened by some to "drinking from a firehose," creates immense opportunities and challenges for libraries. Libraries must anticipate the future implications of this technology, participate in its development, and deploy it to ensure access to the world's information resources.

The infrastructure for the information age is being put in place. Libraries face strategic decisions about their role in the development, deployment, and use of this infrastructure. The emerging infrastructure is much more than computers and communication lines. It is more than the ability to compute at a remote site, send electronic mail to a peer across the country, or move a file from one library to another. The next five years will witness substantial development of the information infrastructure of the network.

In order to provide appropriate leadership, library professionals must have a fundamental understanding of and appreciation for computer networking, from local area networks to the National Research and Education Network (NREN). This presentation addresses these fundamentals, and how they relate to libraries today and in the near future.

Edwin BROWNRIGG Electronic Library Visions and Realities

The electronic library has been a vision desired by many--and rejected by some--since Vannevar Bush coined the term memex to describe an automated, intelligent, personal information system. Variations on this vision have included Ted Nelson's Xanadau, Alan Kay's Dynabook, and Lancaster's "paperless library," with the most recent incarnation being the "Knowledge Navigator" described by John Scully of Apple. But the reality of library service has been less visionary and the leap to the electronic library has eluded universities, publishers, and information technology files.

The Memex Research Institute (MemRI), an independent, nonprofit research and development organization, has created an Electronic Library Program of shared research and development in order to make the collective vision more concrete. The program is working toward the creation of large, indexed publicly available electronic image collections of published documents in academic, special, and public libraries. This strategic

plan is the result of the first stage of the program, which has been an investigation of the information technologies available to support such an effort, the economic parameters of electronic service compared to traditional library operations, and the business and political factors affecting the shift from print distribution to electronic networked access.

The strategic plan envisions a combination of publicly searchable access databases, image (and text) document collections stored on network "file servers," local and remote network access, and an intellectual property management-control system. This combination of technology and information content is defined in this plan as an E-library or E-library collection. Some participating sponsors are already developing projects based on MemRI's recommended directions.

The E-library strategy projected in this plan is a visionary one that can enable major changes and improvements in academic, public, and special library service. This vision is, though, one that can be realized with today's technology. At the same time, it will challenge the political and social structure within which libraries operate: in academic libraries, the traditional emphasis on local collections, extending to accreditation issues; in public libraries, the potential of electronic branch and central libraries fully available to the public; and for special libraries, new opportunities for shared collections and networks.

The environment in which this strategic plan has been developed is, at

the moment, dominated by a sense of library limits. The continued expansion and rapid growth of local academic library collections is now clearly at an end. Corporate libraries, and even law libraries, are faced with operating within a difficult economic climate, as well as with very active competition from commercial information sources. For example, public libraries may be seen as a desirable but not critical municipal service in a time when the budgets of safety and health agencies are being cut back.

Further, libraries in general have a very high labor-to-cost ratio in their budgets, and labor costs are still increasing, notwithstanding automation investments. It is difficult for libraries to obtain capital, startup, or seed funding for innovative activities, and those technology-intensive initiatives that offer the potential of decreased labor costs can provoke the opposition of library staff.

However, libraries have achieved some considerable successes in the past two decades by improving both their service and their credibility within their organizations--and these positive changes have been accomplished mostly with judicious use of information technologies. The advances in computing and information technology have been well-chronicled: the continuing precipitous drop in computing costs, the growth of the Internet and private networks, and the explosive increase in publicly available information databases.

For example, OCLC has become one of the largest computer network

organizations in the world by creating a cooperative cataloging network of more than 6,000 libraries worldwide. On-line public access catalogs now serve millions of users on more than 50,000 dedicated terminals in the United States alone. The University of California MELVYL on-line catalog system has now expanded into an index database reference service and supports more than six million searches a year. And, libraries have become the largest group of customers of CD-ROM publishing technology; more than 30,000 optical media publications such as those offered by InfoTrac and Silver Platter are subscribed to by U.S. libraries.

This march of technology continues and in the next decade will result in further innovations that are extremely difficult to predict. What is clear is that libraries can now go beyond automation of their order files and catalogs to automation of their collections themselves--and it is possible to circumvent the fiscal limitations that appear to obtain today.

This Electronic Library Strategic Plan recommends a paradigm shift in library service, and demonstrates the steps necessary to provide improved library services with limited capacities and operating investments.

SESSION IV-A

Anne KENNEY

The Cornell/Xerox Joint Study in Digital Preservation resulted in the recording of 1,000 brittle books as 600-dpi digital images and the production, on demand, of high-quality and archivally sound paper replacements. The project, which was supported by the Commission on Preservation and Access, also investigated some of the issues surrounding scanning, storing, retrieving, and providing access to digital images in a network environment.

Anne Kenney will focus on some of the issues surrounding direct scanning as identified in the Cornell Xerox Project. Among those to be discussed are: image versus text capture; indexing and access; image-capture capabilities; a comparison to photocopy and microfilm; production and cost analysis; storage formats, protocols, and standards; and the use of this scanning technology for preservation purposes.

The 600-dpi digital images produced in the Cornell Xerox Project proved highly acceptable for creating paper replacements of deteriorating originals. The 1,000 scanned volumes provided an array of image-capture challenges that are common to nineteenth-century printing techniques and embrittled material, and that defy the use of text-conversion processes. These challenges include diminished contrast between text and background, fragile and deteriorated pages, uneven printing, elaborate type faces, faint and bold text adjacency, handwritten text and annotations, nonRoman languages, and a proliferation of illustrated material embedded in text. The latter category included high-frequency and low-frequency halftones, continuous tone photographs, intricate mathematical drawings, maps, etchings, reverse-polarity drawings, and engravings.

The Xerox prototype scanning system provided a number of important features for capturing this diverse material. Technicians used multiple threshold settings, filters, line art and halftone definitions, autosegmentation, windowing, and software-editing programs to optimize image capture. At the same time, this project focused on production. The goal was to make scanning as affordable and acceptable as photocopying and microfilming for preservation reformatting. A time-and-cost study conducted during the last three months of this project confirmed the economic viability of digital scanning, and these findings will be discussed here.

From the outset, the Cornell Xerox Project was predicated on the use of nonproprietary standards and the use of common protocols when standards did not exist. Digital files were created as TIFF images which were compressed prior to storage using Group 4 CCITT compression. The Xerox software is MS DOS based and utilizes off-the shelf programs such as Microsoft Windows and Wang Image Wizard. The digital library is designed to be hardware-independent and to provide interchangeability with other institutions through network connections. Access to the digital files themselves is two-tiered: Bibliographic records for the computer files are created in RLIN and Cornell's local system and access into the actual digital images comprising a book is provided through a document control structure and a networked image file-server, both of which will be described.

The presentation will conclude with a discussion of some of the issues surrounding the use of this technology as a preservation tool (storage, refreshing, backup).

Pamela ANDRE and Judith ZIDAR

The National Agricultural Library (NAL) has had extensive experience with raster scanning of printed materials. Since 1987, the Library has participated in the National Agricultural Text Digitizing Project (NATDP) a cooperative effort between NAL and forty-five land grant university libraries. An overview of the project will be presented, giving its history and NAL's strategy for the future.

An in-depth discussion of NATDP will follow, including a description of the scanning process, from the gathering of the printed materials to the archiving of the electronic pages. The type of equipment required for a stand-alone scanning workstation and the importance of file management software will be discussed. Issues concerning the images themselves will be addressed briefly, such as image format; black and white versus color; gray scale versus dithering; and resolution.

Also described will be a study currently in progress by NAL to evaluate the usefulness of converting microfilm to electronic images in order to improve access. With the cooperation of Tuskegee University, NAL has selected three reels of microfilm from a collection of sixty-seven reels containing the papers, letters, and drawings of George Washington Carver.

The three reels were converted into 3,500 electronic images using a specialized microfilm scanner. The selection, filming, and indexing of this material will be discussed.

Donald WATERS

Project Open Book, the Yale University Library's effort to convert 10,000 books from microfilm to digital imagery, is currently in an advanced state of planning and organization. The Yale Library has selected a major vendor to serve as a partner in the project and as systems integrator. In its proposal, the successful vendor helped isolate areas of risk and uncertainty as well as key issues to be addressed during the life of the project. The Yale Library is now poised to decide what material it will convert to digital image form and to seek funding, initially for the first phase and then for the entire project.

The proposal that Yale accepted for the implementation of Project Open Book will provide at the end of three phases a conversion subsystem, browsing stations distributed on the campus network within the Yale Library, a subsystem for storing 10,000 books at 200 and 600 dots per inch, and network access to the image printers. Pricing for the system implementation assumes the existence of Yale's campus ethernet network and its high-speed image printers, and includes other requisite hardware and software, as well as system integration services. Proposed operating costs include hardware and software maintenance, but do not include estimates for the facilities management of the storage devices and image

servers.

Yale selected its vendor partner in a formal process, partly funded by the Commission for Preservation and Access. Following a request for proposal, the Yale Library selected two vendors as finalists to work with Yale staff to generate a detailed analysis of requirements for Project Open Book. Each vendor used the results of the requirements analysis to generate and submit a formal proposal for the entire project. This competitive process not only enabled the Yale Library to select its primary vendor partner but also revealed much about the state of the imaging industry, about the varying, corporate commitments to the markets for imaging technology, and about the varying organizational dynamics through which major companies are responding to and seeking to develop these markets.

Project Open Book is focused specifically on the conversion of images from microfilm to digital form. The technology for scanning microfilm is readily available but is changing rapidly. In its project requirements, the Yale Library emphasized features of the technology that affect the technical quality of digital image production and the costs of creating and storing the image library: What levels of digital resolution can be achieved by scanning microfilm? How does variation in the quality of microfilm, particularly in film produced to preservation standards, affect the quality of the digital images? What technologies can an operator effectively and economically apply when scanning film to separate two-up images and to control for and correct image imperfections? How can quality control best be integrated into

digitizing work flow that includes document indexing and storage?

The actual and expected uses of digital images--storage, browsing, printing, and OCR--help determine the standards for measuring their quality. Browsing is especially important, but the facilities available for readers to browse image documents is perhaps the weakest aspect of imaging technology and most in need of development. As it defined its requirements, the Yale Library concentrated on some fundamental aspects of usability for image documents: Does the system have sufficient flexibility to handle the full range of document types, including monographs, multi-part and multivolume sets, and serials, as well as manuscript collections? What conventions are necessary to identify a document uniquely for storage and retrieval? Where is the database of record for storing bibliographic information about the image document? How are basic internal structures of documents, such as pagination, made accessible to the reader? How are the image documents physically presented on the screen to the reader?

The Yale Library designed Project Open Book on the assumption that microfilm is more than adequate as a medium for preserving the content of deteriorated library materials. As planning in the project has advanced, it is increasingly clear that the challenge of digital image technology and the key to the success of efforts like Project Open Book is to provide a means of both preserving and improving access to those deteriorated materials.

SESSION IV-B

George THOMA

In the use of electronic imaging for document preservation, there are several issues to consider, such as: ensuring adequate image quality, maintaining substantial conversion rates (through-put), providing unique identification for automated access and retrieval, and accommodating bound volumes and fragile material.

To maintain high image quality, image processing functions are required to correct the deficiencies in the scanned image. Some commercially available systems include these functions, while some do not. The scanned raw image must be processed to correct contrast deficiencies--both poor overall contrast resulting from light print and/or dark background, and variable contrast resulting from stains and bleed-through. Furthermore, the scan density must be adequate to allow legibility of print and sufficient fidelity in the pseudo-halftoned gray material. Borders or page-edge effects must be removed for both compactibility and aesthetics. Page skew must be corrected for aesthetic reasons and to enable accurate character recognition if desired. Compound images consisting of both two-toned text and gray-scale illustrations must be processed appropriately to retain the quality of each.

SESSION IV-C

Jean BARONAS

Standards publications being developed by scientists, engineers, and business managers in Association for Information and Image Management (AIIM) standards committees can be applied to electronic image management (EIM) processes including: document (image) transfer, retrieval and evaluation; optical disk and document scanning; and document design and conversion. When combined with EIM system planning and operations, standards can assist in generating image databases that are interchangeable among a variety of systems. The applications of different approaches for image-tagging, indexing, compression, and transfer often cause uncertainty concerning EIM system compatibility, calibration, performance, and upward compatibility, until standard implementation parameters are established. The AIIM standards that are being developed for these applications can be used to decrease the uncertainty, successfully integrate imaging processes, and promote "open systems." AIIM is an accredited American National Standards Institute (ANSI) standards developer with more than twenty committees comprised of 300 volunteers representing users, vendors, and manufacturers. The standards publications that are developed in these committees have national acceptance and provide the basis for international harmonization in the development of new International Organization for Standardization (ISO) standards.

This presentation describes the development of AIIM's EIM standards and a

new effort at AIIM, a database on standards projects in a wide framework of imaging industries including capture, recording, processing, duplication, distribution, display, evaluation, and preservation. The AIIM Imagery Database will cover imaging standards being developed by many organizations in many different countries. It will contain standards publications' dates, origins, related national and international projects, status, key words, and abstracts. The ANSI Image Technology Standards Board requested that such a database be established, as did the ISO/International Electrotechnical Commission Joint Task Force on Imagery. AIIM will take on the leadership role for the database and coordinate its development with several standards developers.

Patricia BATTIN

Characteristics of standards for digital imagery:

* Nature of digital technology implies continuing volatility.

* Precipitous standard-setting not possible and probably not desirable.

* Standards are a complex issue involving the medium, the hardware, the software, and the technical capacity for reproductive fidelity and clarity.

* The prognosis for reliable archival standards (as defined by librarians) in the foreseeable future is poor.

Significant potential and attractiveness of digital technology as a preservation medium and access mechanism.

Productive use of digital imagery for preservation requires a reconceptualizing of preservation principles in a volatile, standardless world.

Concept of managing continuing access in the digital environment rather than focusing on the permanence of the medium and long-term archival standards developed for the analog world.

Transition period: How long and what to do?

* Redefine "archival."

* Remove the burden of "archival copy" from paper artifacts.

* Use digital technology for storage, develop management strategies for refreshing medium, hardware and software.

* Create acid-free paper copies for transition period backup
until we develop reliable procedures for ensuring continuing
access to digital files.

SESSION IV-D

Stuart WEIBEL The Role of SGML Markup in the CORE Project (6)

The emergence of high-speed telecommunications networks as a basic feature of the scholarly workplace is driving the demand for electronic document delivery. Three distinct categories of electronic publishing/republishing are necessary to support access demands in this emerging environment:

1.) Conversion of paper or microfilm archives to electronic format
2.) Conversion of electronic files to formats tailored to
electronic retrieval and display
3.) Primary electronic publishing (materials for which the
electronic version is the primary format)

OCLC has experimental or product development activities in each of these areas. Among the challenges that lie ahead is the integration of these three types of information stores in coherent distributed systems.

The CORE (Chemistry Online Retrieval Experiment) Project is a model for

the conversion of large text and graphics collections for which electronic typesetting files are available (category 2). The American Chemical Society has made available computer typography files dating from 1980 for its twenty journals. This collection of some 250 journal-years is being converted to an electronic format that will be accessible through several end-user applications.

The use of Standard Generalized Markup Language (SGML) offers the means to capture the structural richness of the original articles in a way that will support a variety of retrieval, navigation, and display options necessary to navigate effectively in very large text databases.

An SGML document consists of text that is marked up with descriptive tags that specify the function of a given element within the document. As a formal language construct, an SGML document can be parsed against a document-type definition (DTD) that unambiguously defines what elements are allowed and where in the document they can (or must) occur. This formalized map of article structure allows the user interface design to be uncoupled from the underlying database system, an important step toward interoperability. Demonstration of this separability is a part of the CORE project, wherein user interface designs born of very different philosophies will access the same database.

NOTES:

(6) The CORE project is a collaboration among Cornell University's Mann Library, Bell Communications Research (Bellcore), the American

Chemical Society (ACS), the Chemical Abstracts Service (CAS), and OCLC.

Michael LESK The CORE Electronic Chemistry Library

A major on-line file of chemical journal literature complete with graphics is being developed to test the usability of fully electronic access to documents, as a joint project of Cornell University, the American Chemical Society, the Chemical Abstracts Service, OCLC, and Bellcore (with additional support from Sun Microsystems, Springer-Verlag, Digital Equipment Corporation, Sony Corporation of America, and Apple Computers). Our file contains the American Chemical Society's on-line journals, supplemented with the graphics from the paper publication. The indexing of the articles from Chemical Abstracts Documents is available in both image and text format, and several different interfaces can be used. Our goals are (1) to assess the effectiveness and acceptability of electronic access to primary journals as compared with paper, and (2) to identify the most desirable functions of the user interface to an electronic system of journals, including in particular a comparison of page-image display with ASCII display interfaces. Early experiments with chemistry students on a variety of tasks suggest that searching tasks are completed much faster with any electronic system than with paper, but that for reading all versions of the articles are roughly equivalent.

Pamela ANDRE and Judith ZIDAR

Text conversion is far more expensive and time-consuming than image capture alone. NAL's experience with optical character recognition (OCR) will be related and compared with the experience of having text rekeyed. What factors affect OCR accuracy? How accurate does full text have to be in order to be useful? How do different users react to imperfect text? These are questions that will be explored. For many, a service bureau may be a better solution than performing the work inhouse; this will also be discussed.

SESSION VI

Marybeth PETERS

Copyright law protects creative works. Protection granted by the law to authors and disseminators of works includes the right to do or authorize the following: reproduce the work, prepare derivative works, distribute the work to the public, and publicly perform or display the work. In addition, copyright owners of sound recordings and computer programs have the right to control rental of their works. These rights are not unlimited; there are a number of exceptions and limitations.

An electronic environment places strains on the copyright system. Copyright owners want to control uses of their work and be paid for any use; the public wants quick and easy access at little or no cost. The marketplace is working in this area. Contracts, guidelines on electronic

use, and collective licensing are in use and being refined.

Issues concerning the ability to change works without detection are more difficult to deal with. Questions concerning the integrity of the work and the status of the changed version under the copyright law are to be addressed. These are public policy issues which require informed dialogue.

*** *** *** ****** *** *** ***

Appendix III: DIRECTORY OF PARTICIPANTS

PRESENTERS:

Pamela Q.J. Andre

Associate Director, Automation

National Agricultural Library

10301 Baltimore Boulevard

Beltsville, MD 20705-2351

Phone: (301) 504-6813

Fax: (301) 504-7473

E-mail: INTERNET: PANDRE@ASRR.ARSUSDA.GOV

Jean Baronas, Senior Manager

Department of Standards and Technology

Association for Information and Image Management (AIIM)

1100 Wayne Avenue, Suite 1100

Silver Spring, MD 20910

Phone: (301) 587-8202

Fax: (301) 587-2711

Patricia Battin, President

The Commission on Preservation and Access

1400 16th Street, N.W.

Suite 740

Washington, DC 20036-2217

Phone: (202) 939-3400

Fax: (202) 939-3407

E-mail: CPA@GWUVM.BITNET

Howard Besser

Centre Canadien d'Architecture

(Canadian Center for Architecture)

1920, rue Baile

Montreal, Quebec H3H 2S6

CANADA

Phone: (514) 939-7001

Fax: (514) 939-7020

E-mail: howard@lis.pitt.edu

Edwin B. Brownrigg, Executive Director

Memex Research Institute

422 Bonita Avenue

Roseville, CA 95678

Phone: (916) 784-2298

Fax: (916) 786-7559

E-mail: BITNET: MEMEX@CALSTATE.2

Eric M. Calaluca, Vice President

Chadwyck-Healey, Inc.

1101 King Street

Alexandria, VA 223l4

Phone: (800) 752-05l5

Fax: (703) 683-7589

James Daly

4015 Deepwood Road

Baltimore, MD 21218-1404

Phone: (410) 235-0763

Ricky Erway, Associate Coordinator

American Memory

Library of Congress

Phone: (202) 707-6233

Fax: (202) 707-3764

Carl Fleischhauer, Coordinator

American Memory

Library of Congress

Phone: (202) 707-6233

Fax: (202) 707-3764

Joanne Freeman

2000 Jefferson Park Avenue, No. 7

Charlottesville, VA 22903

Prosser Gifford

Director for Scholarly Programs

Library of Congress

Phone: (202) 707-1517

Fax: (202) 707-9898

E-mail: pgif@seq1.loc.gov

Jacqueline Hess, Director

National Demonstration Laboratory

for Interactive Information Technologies

Library of Congress

Phone: (202) 707-4157

Fax: (202) 707-2829

Susan Hockey, Director

Center for Electronic Texts in the Humanities (CETH)

Alexander Library

Rutgers University

169 College Avenue

New Brunswick, NJ 08903

Phone: (908) 932-1384

Fax: (908) 932-1386

E-mail: hockey@zodiac.rutgers.edu

William L. Hooton, Vice President

Business & Technical Development

Imaging & Information Systems Group

I-NET

6430 Rockledge Drive, Suite 400

Bethesda, MD 208l7

Phone: (301) 564-6750

Fax: (513) 564-6867

Anne R. Kenney, Associate Director

Department of Preservation and Conservation

701 Olin Library

Cornell University

Ithaca, NY 14853

Phone: (607) 255-6875

Fax: (607) 255-9346

E-mail: LYDY@CORNELLA.BITNET

Ronald L. Larsen

Associate Director for Information Technology

University of Maryland at College Park

Room B0224, McKeldin Library

College Park, MD 20742-7011

Phone: (301) 405-9194

Fax: (301) 314-9865

E-mail: rlarsen@libr.umd.edu

Maria L. Lebron, Managing Editor

The Online Journal of Current Clinical Trials

l333 H Street, N.W.

Washington, DC 20005

Phone: (202) 326-6735

Fax: (202) 842-2868

E-mail: PUBSAAAS@GWUVM.BITNET

Michael Lesk, Executive Director

Computer Science Research

Bell Communications Research, Inc.

Rm 2A-385

445 South Street

Morristown, NJ 07960-l9l0

Phone: (201) 829-4070

Fax: (201) 829-5981

E-mail: lesk@bellcore.com (Internet) or bellcore!lesk (uucp)

Clifford A. Lynch

Director, Library Automation

University of California,

Office of the President

300 Lakeside Drive, 8th Floor

Oakland, CA 94612-3350

Phone: (510) 987-0522

Fax: (510) 839-3573

E-mail: calur@uccmvsa

Avra Michelson

National Archives and Records Administration

NSZ Rm. 14N

7th & Pennsylvania, N.W.

Washington, D.C. 20408

Phone: (202) 501-5544

Fax: (202) 501-5533

E-mail: tmi@cu.nih.gov

Elli Mylonas, Managing Editor

Perseus Project

Department of the Classics

Harvard University

319 Boylston Hall

Cambridge, MA 02138

Phone: (617) 495-9025, (617) 495-0456 (direct)

Fax: (617) 496-8886

E-mail: Elli@IKAROS.Harvard.EDU or elli@wjh12.harvard.edu

David Woodley Packard

Packard Humanities Institute

300 Second Street, Suite 201

Los Altos, CA 94002

Phone: (415) 948-0150 (PHI)

Fax: (415) 948-5793

Lynne K. Personius, Assistant Director

Cornell Information Technologies for

Scholarly Information Sources

502 Olin Library

Cornell University

Ithaca, NY 14853

Phone: (607) 255-3393

Fax: (607) 255-9346

E-mail: JRN@CORNELLC.BITNET

Marybeth Peters

Policy Planning Adviser to the

Register of Copyrights

Library of Congress

Office LM 403

Phone: (202) 707-8350

Fax: (202) 707-8366

C. Michael Sperberg-McQueen

Editor, Text Encoding Initiative

Computer Center (M/C 135)

University of Illinois at Chicago

Box 6998

Chicago, IL 60680

Phone: (312) 413-0317

Fax: (312) 996-6834

E-mail: u35395@uicvm..cc.uic.edu or u35395@uicvm.bitnet

George R. Thoma, Chief

Communications Engineering Branch

National Library of Medicine

8600 Rockville Pike

Bethesda, MD 20894

Phone: (301) 496-4496

Fax: (301) 402-0341

E-mail: thoma@lhc.nlm.nih.gov

Dorothy Twohig, Editor

The Papers of George Washington

504 Alderman Library

University of Virginia

Charlottesville, VA 22903-2498

Phone: (804) 924-0523

Fax: (804) 924-4337

Susan H. Veccia, Team leader

American Memory, User Evaluation

Library of Congress

American Memory Evaluation Project

Phone: (202) 707-9104

Fax: (202) 707-3764

E-mail: svec@seq1.loc.gov

Donald J. Waters, Head

Systems Office

Yale University Library

New Haven, CT 06520

Phone: (203) 432-4889

Fax: (203) 432-7231

E-mail: DWATERS@YALEVM.BITNET or DWATERS@YALEVM.YCC.YALE.EDU

Stuart Weibel, Senior Research Scientist

OCLC

6565 Frantz Road

Dublin, OH 43017

Phone: (614) 764-608l

Fax: (614) 764-2344

E-mail: INTERNET: Stu@rsch.oclc.org

Robert G. Zich

Special Assistant to the Associate Librarian

for Special Projects

Library of Congress

Phone: (202) 707-6233

Fax: (202) 707-3764

E-mail: rzic@seq1.loc.gov

Judith A. Zidar, Coordinator

National Agricultural Text Digitizing Program

Information Systems Division

National Agricultural Library

10301 Baltimore Boulevard

Beltsville, MD 20705-2351

Phone: (301) 504-6813 or 504-5853

Fax: (301) 504-7473

E-mail: INTERNET: JZIDAR@ASRR.ARSUSDA.GOV

OBSERVERS:

Helen Aguera, Program Officer

Division of Research

Room 318

National Endowment for the Humanities

1100 Pennsylvania Avenue, N.W.

Washington, D.C. 20506

Phone: (202) 786-0358

Fax: (202) 786-0243

M. Ellyn Blanton, Deputy Director

National Demonstration Laboratory

for Interactive Information Technologies

Library of Congress

Phone: (202) 707-4157

Fax: (202) 707-2829

Charles M. Dollar

National Archives and Records Administration

NSZ Rm. 14N

7th & Pennsylvania, N.W.

Washington, DC 20408

Phone: (202) 501-5532

Fax: (202) 501-5512

Jeffrey Field, Deputy to the Director

Division of Preservation and Access

Room 802

National Endowment for the Humanities

1100 Pennsylvania Avenue, N.W.

Washington, DC 20506

Phone: (202) 786-0570

Fax: (202) 786-0243

Lorrin Garson

American Chemical Society

Research and Development Department

1155 16th Street, N.W.

Washington, D.C. 20036

Phone: (202) 872-4541

Fax: E-mail: INTERNET: LRG96@ACS.ORG

William M. Holmes, Jr.

National Archives and Records Administration

NSZ Rm. 14N

7th & Pennsylvania, N.W.

Washington, DC 20408

Phone: (202) 501-5540

Fax: (202) 501-5512

E-mail: WHOLMES@AMERICAN.EDU

Sperling Martin

Information Resource Management

20030 Doolittle Street

Gaithersburg, MD 20879

Phone: (301) 924-1803

Michael Neuman, Director

The Center for Text and Technology

Academic Computing Center

238 Reiss Science Building

Georgetown University

Washington, DC 20057

Phone: (202) 687-6096

Fax: (202) 687-6003

E-mail: neuman@guvax.bitnet, neuman@guvax.georgetown.edu

Barbara Paulson, Program Officer

Division of Preservation and Access

Room 802

National Endowment for the Humanities

1100 Pennsylvania Avenue, N.W.

Washington, DC 20506

Phone: (202) 786-0577

Fax: (202) 786-0243

Allen H. Renear

Senior Academic Planning Analyst

Brown University Computing and Information Services

115 Waterman Street

Campus Box 1885

Providence, R.I. 02912

Phone: (401) 863-7312

Fax: (401) 863-7329

E-mail: BITNET: Allen@BROWNVM or

INTERNET: Allen@brownvm.brown.edu

Susan M. Severtson, President

Chadwyck-Healey, Inc.

1101 King Street

Alexandria, VA 223l4

Phone: (800) 752-05l5

Fax: (703) 683-7589

Frank Withrow

U.S. Department of Education

555 New Jersey Avenue, N.W.

Washington, DC 20208-5644

Phone: (202) 219-2200

Fax: (202) 219-2106

(LC STAFF)

Linda L. Arret

Machine-Readable Collections Reading Room LJ 132

(202) 707-1490

John D. Byrum, Jr.

Descriptive Cataloging Division LM 540

(202) 707-5194

Mary Jane Cavallo

Science and Technology Division LA 5210

(202) 707-1219

Susan Thea David

Congressional Research Service LM 226

(202) 707-7169

Robert Dierker

Senior Adviser for Multimedia Activities LM 608

(202) 707-6151

William W. Ellis

Associate Librarian for Science and Technology LM 611

(202) 707-6928

Ronald Gephart

Manuscript Division LM 102

(202) 707-5097

James Graber

Information Technology Services LM G51

(202) 707-9628

Rich Greenfield

American Memory LM 603

(202) 707-6233

Rebecca Guenther

Network Development LM 639

(202) 707-5092

Kenneth E. Harris

Preservation LM G21

(202) 707-5213

Staley Hitchcock

Manuscript Division LM 102

(202) 707-5383

Bohdan Kantor

Office of Special Projects LM 612

(202) 707-0180

John W. Kimball, Jr

Machine-Readable Collections Reading Room LJ 132

(202) 707-6560

Basil Manns

Information Technology Services LM G51

(202) 707-8345

Sally Hart McCallum

Network Development LM 639

(202) 707-6237

Dana J. Pratt

Publishing Office LM 602

(202) 707-6027

Jane Riefenhauser

American Memory LM 603

(202) 707-6233

William Z. Schenck

Collections Development LM 650

(202) 707-7706

Chandru J. Shahani

Preservation Research and Testing Office (R&T) LM G38

(202) 707-5607

William J. Sittig

Collections Development LM 650

(202) 707-7050

Paul Smith

Manuscript Division LM 102

(202) 707-5097

James L. Stevens

Information Technology Services LM G51

(202) 707-9688

Karen Stuart

Manuscript Division LM 130

(202) 707-5389

Tamara Swora

Preservation Microfilming Office LM G05

(202) 707-6293

Sarah Thomas

Collections Cataloging LM 642

(202) 707-5333

END

****

Note: This file has been edited for use on computer networks. This

editing required the removal of diacritics, underlining, and fonts such

as italics and bold.

kde 11/92

[A few of the italics (when used for emphasis) were replaced by CAPS mh]