NINCH SYMPOSIUM: April 8, 2003, New York City

REPORT

The Price of Digitization: New Cost Models for Cultural and Educational Institutions

A Digitization Symposium Presented by NINCH and Innodata

Co-sponsored by The New York Public Library and New York University

---

Report by Lorna Hughes
New York University

Also available in: MSWord

Also available on the website of the Canadian Heritage Information Network

Summary Report by Michael Lesk

Agenda | Speaker Bios | Resources |

INTRODUCTIONS

Peter B. Kaufman *Innodata*

Peter Kaufman, Director of Strategic Initiatives at Innodata, opened the meeting by welcoming the 260 participants from libraries, museums, and universities, as well as staff from Innodata and other digitization vendors. In his brief remarks, Kaufman acknowledged the sponsors of the event: Innodata, New York University, New York Public Library, and NINCH. Specific thanks were given to Innodata's Chairman and marketing department, and to David Green of NINCH

Describing the services offered by Innodata, such as training, professional consultancy or responding to requests for services, Kaufman expressed his hope that this meeting would present an opportunity for people in the same sector to share knowledge and information. He also hoped it would provide an opportunity to focus on and possibly enhance the *NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials*, with particular emphasis on the Guide's sections on "Cost Models", "Project Planning", and "Working Together". He outlined a vision that the next edition of the Guide could include a suite of representative RFPs, completed proposals, outlines and mini-planning guides. Both vendors and cultural heritage organizations are actively involved in developing such materials, and coordination across the sectors could lead to the emergence of useful standards on common goals and terminology. Noting that this meeting had attracted such a substantial enrollment, clearly reflecting a great deal of interest in this topic and a desire for information throughout the community, Kaufman announced that the event would be repeated at new locations on the West Coast, the Midwest and the Southeast in 2003 and 2004.

Kaufman spoke briefly about the genesis of the idea for the meeting in an article by Margaret Hedstrom, "The Digital Preservation Research Agenda," published in *The State of Digital Preservation: An International Perspective* (CLIR, 2002). There, Hedstrom states that "the challenge of developing economic models for the value and costs of archiving over the long term deserves an entire meeting or conference."

Noting that these concerns affect the cultural heritage community as a whole, Kaufman emphasized the urgency of developing standard practices for preservation and access. Virtually all libraries and museums maintain collections of one-of-a-kind printed and manuscript materials that present significant challenges for preservation and access. Some materials are in great demand, but, because of their value and condition, are endangered by unrestricted use. Many more remain inaccessible to all but the most intrepid researchers because of outdated and/or inadequate descriptions and finding aids. The information contained in archival and local history collections defies most standard library classification systems for several reasons: they are difficult to categorize; their principal value is their uniqueness; and the most effective way to describe documents is to show them (and since historic documents have artifactual as well as

informational value, direct visual contact is usually important to the researcher). For the library, then, a priceless community legacy can become an administrative albatross. The result has been that irreplaceable material deteriorates, sizable sections of library collections are underutilized, and potent historical information sits inaccessible to scholars, educators, community leaders, and the general public.

Kaufman emphasized the incredible ubiquity of activity and concerns about digitization and electronic media. Publishers, universities, politicians and cultural heritage organizations are all thinking about digitization issues and he cited just a few of the major initiatives: The Library of Congress National Digital Information Infrastructure and Preservation Program, the White House Office of Science and Technology Policy, the programs of the National Archives and Records Administration, The National Library of Medicine, the National Agricultural Library and NASA. The Library of Congress receives over 2 million requests a day for digital files, compared to 2 million requests per year for items to be delivered to readers in its rooms, and, according to a recent report, there are now 50 million historical documents posted on the web by the National Archives alone (see http://www.archives.gov/aad/).

Kaufman concluded by calling for a formal forum of exchange to look at the lessons of market discipline learned by the commercial sector: not business models, *per se*, but lessons from business. Such a forum would create a framework for what Clifford Lynch of the Coalition for Networked Information has called “federating”, that is, developing a “fruitful” area for exploration and innovation.

Kaufman hoped this meeting would begin to achieve such an objective.

David Green *NINCH*

David Green, Executive Director of NINCH, expressed his thanks to Vincent Doogan of New York University, and to Heike Kordish and Jan Brown at the New York Public Library for their sponsorship of the meeting, as well as to Innodata, the first member of NINCH’s corporate council and co-organizer of the symposium. Green also acknowledged the remarkable collection of speakers who had agreed to donate their services.

Green cited several NINCH programs that are actively promoting the type of community called for by Peter Kaufman, including the exchange of information via NINCH-announce; nationwide discussions of copyright through the Copyright Town Meetings; and developing interdisciplinary collaborations through the Computer Science and the

Humanities initiative). Green also mentioned the several projects which fall under the auspices of NINCH's "tools for today" series, including the International Database of Digital Humanities Projects. Of all its work, Green commented that he thought the hallmark project is the *NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials*. International in scope, based on empirical interviews with over 30 major digitization sites, and directed by a NINCH Working Group drawn from museums, libraries, archives, scholars and teachers, visual resources and humanities computing services, the Guide both in its process and in its evolving product was an effort to draw in the different expertise and experience of different kinds of institutions all working broadly for the same goal of an interoperable, sustainable body of rich cultural materials in digital form. He outlined plans for its further evolution to reflect changes in the field, especially in the topic at hand, the pricing and costing of digitization.

As Green himself was leaving NINCH, he indicated that he would be continuing the broad mission of networking cultural resources as he had done while heading the coalition.

Addressing the issues of preservation and access, Green singled out for attention the vision of the Andrew W. Mellon Foundation that has spearheaded thinking about digitization for many years. He then introduced Don Waters, the Foundation's Program Officer for Scholarly Communications, who delivered the keynote address.

Donald Waters, *The Economics of Digitizing Library and Other Cultural Materials: Perspective from the Mellon Foundation.*

Waters noted that the theme of this meeting identified a set of issues that had been part of a broader set of concerns at the Andrew W. Mellon Foundation for the better part of a decade, and that a reflection on the grant making activities of the Foundation afforded him an opportunity to frame the discussions in a way that he hoped would be helpful to the audience.

Waters began with definitions of some key concepts: "costs" are the financial or other obligations incurred in the course of producing goods or services. Costs can be indirect or direct; and can be internal or external to the project at hand. Costs will vary in size, and are not self-evident. Indirect costs in particular can be hard to measure, as they are defined by institutional practices and metrics that are not transparent, but they cannot be ignored. Waters cited electronic journals as an example of the shifting paradigms of delivery of resources by libraries. Libraries now rent, rather than purchase serials. The costs of renting vs. buying journals are very different – cost related to buying serials includes the cost of storing, shelving, retrieving and cataloguing the materials, as well as

costs related to the physical storage of the content: the costs of building libraries; the cost of power for heat, light and air conditioning. These are indirect cost to the library acquisitions budget, and to a certain degree to the library itself. The shift to renting electronic content has reduced the costs of maintaining the physical materials, but has increased the cost of preserving the content. Who is paying or is willing to pay to insure against the massive loss of digitized information?

Waters clarified that “price” and “cost” are not the same thing: “Price” is the amount paid by a customer for a good or service and is set in the marketplace. Price may be only tangentially related to cost, and the difference between price and cost defines profit (or loss). Waters referred to the particular issues facing nonprofit organizations in charging prices that meet their costs.

He emphasized that the process of “digitization” must be understood in order to accurately understand cost issues. As the field matures, we realize that digitization is not a uniform process, and that digital interoperability is neither simple nor straightforward. In examining this question, Waters drew parallels with the history of publishing. Citing Adrian Johns, *The Nature of the Book: Print and Knowledge in the Making* (Chicago, 1998), Waters noted that print did not emerge casually, but as the result of laborious processes.

There was very little trust in the print medium when it was first developed – it was seen as unstable and subject to piracy and fraudulent copying. Authenticity was hard to guarantee: indeed, the term “piracy” was first used by John Fell, Bishop of Oxford, to describe certain pernicious practices of early printers and booksellers. A “pirate” was someone who participated in the “unauthorized reprinting of a title recognized to belong to someone else.” “Stationers” eventually emerged as the trusted practitioners who were placed in charge of various aspects of publishing – practices we would now recognize as printing, publishing, editing, and bookselling. Stationers worked out the conventional practices of making books, and thus made printing a viable economic enterprise with the elaborate complexity of producing a book eventually invisible to all but the practitioners in the trade.

Waters likened these stationers to “digitizers, who similarly have had to define and disentangle the various practices associated with digitization, including an examination and careful definition of various tasks and costs involved. In particular, Waters singled out activities such the workshops organized by Kate Wittenberg at Columbia for young scholars who have won a Gutenberg prize to turn their thesis into an e-book. The workshops organize the various processes in the production of an e-book and try to regularize and normalize that production process. He cited the *NINCH Guide to Good Practice* as an especially noteworthy effort to identify and codify current conventions. Such initiatives have given us the power to imagine the real costs of such initiatives. Digitizing the Library of Congress is still prohibitively expensive, but at least we can now

make a reasonable estimate of how much such a project would cost, based on the experiences of projects like the University of Michigan's Making of America project. The care and precision with which digitizing costs are being measured by institutions and projects such as Michigan, Virginia, JSTOR, ARTstor, the Library of Congress, and other institutions that have undertaken large scale digitization projects, are creating an economic discipline that focuses on areas of high cost and results in significant market pressure systematically to reduce those costs as barriers to massive digitization.

Waters identified three cost barriers to digitization. Careful attention to these barriers is critical for jump-starting the dynamic of cost-savings.

1. Technology and workflow costs.
A better understanding of good practices has created a more efficient production workflow, and Waters pointed to some useful innovations (such as those developed and refined by Luna Imaging on large-scale projects for the Museum of Modern Art and the New York Public Library) in developing metadata to track workflow, and in quality control. In addition, technology development paths that result in lower costs have now been clearly identified for various formats, including the capture and markup of text, and OCR. It is possible to make informed choices about the possible tradeoffs related to cost and quality. Processes for digitizing sound and video are much less well developed, but measurable standards are emerging.

2. Intellectual property costs.
The temptation is either to despair of the cost and abandon digitization, or to try to operate under the radar of the "copyright police". Slide digitization projects often apply the latter approach, by restricting access to campus machines, or even to registered students of a particular class. Such initiatives avoid lawsuits, but result in costly duplication of effort across many campuses. Projects attempting to address this issue include ARTstor, JSTOR, CIAO, ACLS's History-E project, the BiblioVault project at the University of Chicago, and the Electronic Enlightenment at Oxford University. Such projects demonstrate that communities of users and publishers can find ways to create the trust and goodwill needed to overcome the costly barriers of copyright and create highly useful digitized collections of research and educational materials.

3. Institutional costs and variables.
The organizational variables that affect decisions about how to approach technology or intellectual property costs factors are rarely recognized or analyzed. There is a need for institutions to be able to define and defend their choices related to digitization in terms of their institutional mission of teaching and research, and to avoid the distraction of commercializing their products. Furthermore, within an institutional context, such clarity of mission will allow the costs of digitization to be offset by economies of scale. For example, purchasing JSTOR means that an institution will not have to incur the cost of new shelf space. On aggregate, these savings could be enormous. Such savings will

come out of different parts of the overall institutional budget, but if they could be captured, these savings would provide a massive fund for further digitization.

In deciding whether digitized resources are worth the cost, institutions need to think in broad terms to take account of all elements of the financial equation: including the long-term implications for building plans, capital costs, and maintenance. In a digital world, a broader institutional perspective needs to be applied to resource allocation decisions. Such major institutional lessons cannot be learned if digitization is tucked away in relatively small digital production departments within a university library. Presidents, provosts, deans, scholars, librarians, and technologists together must find ways within the larger academic community for their institutions to work together to realize the extraordinary economies of scale that are possible, and foundations like Mellon should not be seen as the “deep pockets” to which they turn to cover the huge costs of digitizing, but as catalysts in the necessary effort to establish these new modes of cooperation. Incentives for such collaborations should include the advancement of the academic mission of teaching and research. Electronic resources should facilitate scholarship, enabling primary evidence to be found across institutions by custodial paradigms replicating the serendipity of browsing the library stacks. The demand for such resources should be the opposite side of the economic equation to “costs”. From demand borne of real need follows the income streams that create sustainability, whether it is in the form of contributions, user fees, or base budget support from the home institution.

Waters concluded by mentioning the Society for the Diffusion of Useful Knowledge, a nineteenth-century example of a project to provide cheap books for the working class that was attacked either for distributing dangerous ideas or for distributing no ideas at all. In the ensuing debate about costs and quality, the Society lost sight of its ultimate objective: meeting demand for useful knowledge.

Back to top

CASE STUDIES: CALCULATING PRODUCTION COSTS

Introducing the panel, Peter Kaufman hoped discussions of case studies would enable improved approaches to scholarly collaboration. He also hoped that the discussions at the symposium would dispel the image that pricing for digitization was undertaken in secrecy. Kaufman hoped that developing repositories of such case studies and opportunities for exchanges of information make the pricing process more transparent in future.

Maria Bonn ***Economies of Scale: Lessons Learned from the Making of America IV Project.***

For Presentation Slides: see Powerpoint

Maria Bonn presented an overview of a stacks-driven digital imaging project (part of the larger Making of America project) aimed at preserving and increasing access to embrittled 19th-century American volumes. Instead of microfilming the documents as might have been done a few years ago, the project scanned them using bitonal imaging followed by optical character recognition. The files were then put into an online access system developed at the University of Michigan. There were two phases of the Making of America (MoA) project at Michigan: the 1996 Making of America I consisted of 1,500 volumes, and the 2000 Making of America IV digitized 8,500 volumes. As part of MoA, the materials are freely available to the public, with a reprint service available. Usage of the materials is very high – over a million users accessed the materials January - February, 2003, which is especially significant in light of the fact that the originals had not been in circulation at the library.

As part of the Making of America IV project, The Mellon Foundation funded a study on the costs and methods of using digital technologies for preserving and deploying monographic materials. This study was an attempt to collect, analyze and report data on the costs of all significant phases of digitization of ordinary books. It was used as benchmark to evaluate other digitization proposals.

An article by Bonn describing the project, entitled "Benchmarking Conversion Costs: A Report from the Making of America IV Project", was published in *RLG DigiNews* (October 2001) and a fuller review of costs and methods is available in the report to the Mellon Foundation, *Assessing the Costs of Conversion* (See Resources Page)

The assumptions behind the cost study were:

- That selection costs were negligible if automatic processes were used;
- That the cost per page is the most reliable cost unit;
- That the existence of some institutional infrastructure is key; and
- That staff can and should multi-task, as this will keep them engaged in what can be rather tedious, repetitive tasks.

Bonn highlighted a few *caveats* to adapting this data to other projects: the level of existing infrastructure will vary, as will local practices and labor markets. Furthermore, she pointed out that this data is now over two years old, so the actual numbers would be different today.

The selection criteria for the materials was automatic, making it easier and cheaper to proceed: the project was limited to embrittled monographs (including pamphlets) published between 1850-1876; the materials had to be in US editions, and were all in

English; and they were stored in a remote shelving facility. Despite these fairly broad criteria, materials still had to be reviewed for significant illustrations, to ensure they were not valuable first editions, were signed or contained materials by notable authors. Bindings also had to be examined.

Data collection and analysis for the study included the following components:

- Three sets of time and performance studies spread over the duration of the project;
- An analysis of salary and special equipment costs;
- Use of established rates for scanning and OCR; and
- Cost snapshots at different points in the project.

Activities related to digitization that were tracked included: the retrieval of volumes from storage; charging out of volumes; identification, collation and repair; disbanding and removal of covers; packing and shipping to the vendor; scanning and CD burning; metadata creation; quality control; and OCR and SGML generation.

Detailed costs can be seen in Bonn's presentation slides, including a detailed cost determination. These costs break down to 20-27 cents per page: 13 cents for scanning and the rest for overhead, selection and processing. However, Bonn pointed out that the real total costs per page will vary – this survey was based on a hypothetical "most productive" month, when there was the greatest efficiency in preparation and when OCR and scanning staff were all at their most efficient.

Several questions were raised by the study, which Bonn indicated are still open issues:

- When is the best time to repair and replace materials?
- Should this be factored into the workflow?
- How can a balance be struck between usability, cost and the best representation of the artifact? (For example, should blank pages at the beginning and end of each volume be included?)
- After digitization, should the original volumes be kept? For how long?
- And what is the best way to collaborate with peers on cost-effective physical and virtual preservation and access of digitally reformatted volumes?

Although several questions remain unanswered, there are some important lessons to be learned from the project, including:

- Affection for the artifact can slow production – stopping the digitization process to read the book is not a good idea.
- A higher volume and larger production staff can bring down costs.

- Exceptions always increase costs, such as taking the time to digitize the occasional image, or to handle special preservation concerns.
- For projects, the ramp-up is the hardest part – and the most expensive in terms of "cost per image."
- Grant-funded projects probably can not keep costs consistently low - they will lose efficiencies of scale. This can be improved with collaboration.
- Overall, it was the volume of this project that was the key to keeping costs low.

In questions following the talk, it was clarified that on this project vendors did 100% of the quality control, with the project staff assessing a five percent sample of the materials. Four full-time staff were assigned to the project, as many as twenty other staff members contributed percentages of their time at various stages of the project.

Back to top

Nancy Harm *Luna Imaging: A Manufacturing Model*

Luna Imaging is a California-based image digitization company, specializing in digitizing visual collections, that has developed a project-based model incorporating best practices into the digitization workflow. Harm presented an overview of Luna's services, processes and workflows, and described some of the advantages to be offered by working with an experienced vendor rather than doing in-house digitization.

The "Luna Process" is controlled by a carefully structured web-based tracking system that manages roles and assignments. The phases of the process were described as:

1. Inventory and receipt of deliverables
2. Image Capture
3. Phase One Edit: Color Balance & "Dust Bust"
4. Phase Two Edit: Cropping / Sizing - Derivative Creation
5. Batch and Write to Media, Update Management Data, Initial QC
6. Final QC – Shipping

Harms described the advantages of partnering with Luna lying in its:

- skilled image technicians
- high-end equipment
- experience working on many projects
- working to deadlines

- quantifiable results
- proven quality of service.

Collaboration, she said, can help define a project by focusing on the resources and opportunities available. Luna can bring to bear its experience of working on many projects, including the wisdom gathered from mistakes.

Harms listed some of the questions that will help the project manager to assess whether or not digitization should be outsourced will include:

- *Staff*: Will the project be based in-house with current resources, or with a new team? Do you have a trained staff? Are you able to bring in new staff to start up the project?
- *Timeline:* Are there critical deadlines?
- *Workspace:* Do you have the physical space for these tasks or is the space best used for other activities?
- *Equipment:* Do you have the right mix? Can you capitalize upon these investments through sharing the equipment with other projects?

Defining costs is a process that combines the client's requirements and Luna's in-house capabilities. There will be a wide range of variability, depending on the quality required, the type of materials, and the time necessary. Costs quoted can range from $4 for a 35mm slide to $60 for a larger item. However, some guidelines on cost savings were presented. These include: having accurate data; organized materials; high volume; good communication; consistency, and clearly defined project goals.

Harm emphasized that Luna and its clients have an equal partnership, with complimentary roles. Luna can use its experience to help the client identify priorities and the scale of the project. It can identify and undertake areas of work that the institution is unable to do in-house, and can assess potential roadblocks to project success. Luna can identify workflows and procedures not related to digitization, such as ensuring that the creation and editing of catalog data and tracking entries precedes the start of production. Luna can assign project managers to core activities.

The client's role is to direct Luna's activities, to assess, plan and schedule to projects, and to provide the final quality control. The client will monitor progress and accountability. To illustrate these issues, Harm referenced some of the projects Luna is working on with clients from the cultural heritage community, including the Brooklyn Public Library's Brooklyn Collection; the Getty Research Institute's Tapestry Project; Indiana University's Cushman Archive; the Museum of Modern Art's Digital Design Collection; and Yale University's Beinecke Library

In conclusion, Harm observed that Luna's motto in determining true costs is to invest in quality to recoup the costs over the life of the image. She also noted that it is difficult for institutions to effectively compare output and services from a range of vendors without generalized points of comparison such as bit-depth, resulting file size, and associated vendor services. There is a constant need for more information.

Back to top

Dan Pence ***Ten Ways to Spend $100,000 on Digitization***

See Presentation slides as Powerpoint

In a presentation that was extremely detailed and open about his company's costing processes, Dan Pence described the work carried out by the Systems Integration Group (SIG). SIG is a small, privately held company with 170 employees. Its primary client base is the federal government, and clients include the U.S. Coast Guard; the U.S. Department of Agriculture; the U.S. Department of State; the U.S. Department of Treasury; the Pension Benefit Guaranty Carp; the Environmental Protection Agency and the National Park Service.

In terms of cultural heritage projects, SIG has a 10-year relationship with the Library of Congress, including work done for the American Memory Project and the National Digital Library Program (total metrics of the LC work include1.5+ million pages digitized and 1.6 + billion characters of text conversion). Other Cultural Heritage projects include digital library projects for the National Agricultural Library (6 years); New York University; New York Botanical Garden; the Newseum, and the Smithsonian.

Pence illustrated some core advantages to outsourcing:

1. Knowledgeable and committed staff will already be in place, meaning that there will be minimal start-up delay and disruption and a high level of productivity from the beginning, and no hiring or training concerns
2. Process tools are in already in place, so there will be no lost or lag time to deliver high quality images, and production staff are already familiar with format and handling
3. No need for capital investment – no equipment maintenance headaches, and no technology refreshment/upgrade issues

He then outlined some guidelines for working with vendors. Prior to starting the project,

the vendor will need to know:

- General Nature of the Project
- General description of the materials to be digitized
- Ultimate objectives for the project, i.e.: how will the products of digitization be used?
- Place of performance – on/off site
- Anticipated or desired schedule

A vendor will also need to know the characteristics of the collection, for example:

- What constitutes an item or a document?
- How many items are there to be digitized?
- Are the items bound or unbound?
- What are the page dimensions?
- When will the documents be available?
- What are the handling specifications?
- What are the insurance requirements?

Specifications for the products of digitization will also have to be negotiated, metrics such as: the digital imaging resolution – "dpi"; tonality (bitonal, grayscale, or color); the desired file format – TIFF, JPEG, PDF, etc., and whether or not compression is acceptable; the directory and file naming requirements; indexing or metadata requirements; and the delivery medium (e.g., CD-ROM, WWW) and quantity.

Pence outlined some important factors that will increase the cost of digital imaging, including image file size, special handling requirements, and interruptions in workflow. The factors that determine image file size are: (1) bit depth (e.g., bitonal/grayscale/color); resolution (e.g., 300 dpi/600 dpi); and page size (i.e., small/large). Factors that impact the cost of handling are: (1) whether or not items are bound and (2) whether or not items are fragile. Interruptions in workflow may occur when (1) the scanning operation outpaces the document preparation operation or (2) the collection is characterized by a small number of pages per item, resulting in high administrative, indexing, or tracking costs relative to the total number of pages scanned.

He then presented his "Chinese menu", a standard-price menu that quickly illustrated how many variables are to be found in attempting to develop cost metrics for digitization:

Category	Choices
Bound/unbound	2
Page size (8.5x11, 11x17, 17x22)	3

Scanning resolution (300, 400, 600)	3
Scanning bit depth (1, 8, 24)	3
Handling (fragile/non-fragile)	2
Place of performance (on/off site)	2
Possible combinations	216

These choices also ignore the added factor of possible discounts; price breaks for quantity, etc. As reality sets in, it becomes apparent that a standard price schedule is only the starting point for a cost estimate. Every digitization project has a unique profile and must be priced individually.

A vendor's fixed unit price is determined by the following variables:

- Effective number of pages scanned per day
- Direct cost of scanning labor
- Direct cost of post-processing & QA labor
- Amortization of the cost of equipment
- Overhead and profit

The number of pages that can be scanned per day will be a function of:

- Document page size and binding
- Handling specifications
- Digital image format specifications
- Scanning technology availability

In his slides, Pence presented a hypothetical case study developing a quote for digitizing scientific volumes from the nineteenth century. The project consists of 6,300 pages, including 311 color maps. He used this to extract a cost figure of $12 per page for text pages, and $12 per page for maps, at a total of $16,332.

One metric that cannot be quantified is that vendors will take responsibility of a large number of potential project risk factors:

- Equipment will be fully utilized over 3 years
- Equipment failure will be minimal
- Software upgrades will not cause problems
- Supply of material will not be interrupted
- Employees will show up for work
- Employees will maintain productivity

- Employees will handle material carefully
- Image quality will meet or exceed requirements, resulting in little or no rework.

Pence then sated the curiosity of the audience by answering the question in his title: The 10 ways he suggested to spend $100,000 on digitization were:

1. Purchase $100,000 of equipment and software and hope that you have the budget for staff next year, or
    1. Pay a vendor $16,332 and use the remainder for …
2. Adding to your original source collection
3. Digitizing more of your collection
4. Preserving fragile source materials
5. Adding permanent staff
6. Funding internships
7. Training existing staff
8. Enhancing your web site
9. Installing security equipment
10. Investing in digital library capabilities

In the question session, it was pointed out that even cheaper digitization can be accomplished by sending materials overseas for digitization (an approach taken by Innodata and exemplified by the Carnegie Mellon/Internet Archive Million Book Project). Costs can be reduced greatly if the project will accept disbinding, shipping overseas, and lower quality.

Another questioner suggested that the type of costing worksheets illustrated in this session might make a useful addition to subsequent editions of the *NINCH Guide to Good Practice*.

Back to top

Peter B. Kaufman *Digitizing History: University Presses and Libraries*

Kaufman referenced Clifford Lynch's analysis of the typical limits of scholarly publishing in the genres in which it is disseminated, and Lynch's vision for a time and place where the institutional repository can serve as a complement or supplement—not a substitute—to traditional scholarly publication. Such a repository could capture and disseminate learning and teaching material, symposia, performances and related documentation of the life of universities. Many collaborators could be involved in such

initiatives: libraries could join forces with local governments, historical societies, museums and archives, and members of the community, and public broadcasting might play a role. Kaufman also mentioned several new licensing and media opportunities available to libraries, universities, and museums that are likely to generate revenue.

Kaufman then presented three case studies of university projects that have used Innodata's expertise and that illustrated such shifting paradigms of scholarship and communication:

The University of Virginia requested support in transforming a rare and extensive medical history collection, the Philip S. Hench Walter Reed Yellow Fever Collection. He spent over fifteen years accumulating thousands of documents, photographs, miscellaneous printed materials, and artifacts to decipher the actual events involved in the U.S. Army Yellow Fever Commission work in Cuba at the turn of the 20th century. The archive consists of some 30,000 pages of manuscripts in English, Spanish, and French; technical pamphlets and books; newspapers; photographs; artifacts; and research. These were converted into digital form (XML) for the purposes of preservation and access.

The University of Illinois Press asked Innodata to convert digital files from the leading historical journals in XYAscii, Adobe Frame Maker, Pagemaker, and Quark file formats into a kind of TEI Lite. Journals included: the *American Historical Review;* the *Journal of American History; Labour History; Law and History Review; The History Teacher; Western Historical Quarterly;* and the *William and Mary Quarterly*.

At the University of North Carolina, an ongoing initiative involves the UNC libraries, press, school of information, school of journalism, faculty, the UNC-TV station, and WUNC, and is developing a service center on the model of Cornell's or Reed College's.

Back to top

PRESERVATION COSTS

Stephen Chapman Counting the Costs of Digital Preservation: Is Repository Storage Affordable?

See Chapman's Presentation Slides: in Powerpoint
See Chapman's article on this research in the Journal of Digital Information

Chapman's talk was a contribution to the debate on the costs of digital preservation,

which must be considered at the outset of any digitization project. His presentation examined the costs of paper versus digital repositories for managed, long-term storage of library materials, based on case studies of OCLC's Digital Archive and the Harvard Depository.

Chapman began by addressing the benefits of such long-term storage repositories. Benefits are related to risk management: long-term storage in a managed repository will provide an "insurance" against the following risk factors:

- obsolescence
- inadvertent loss or theft.
- technology changes
- the evolution of user expectations, as usage patterns change.

The costs of preservation are contextual – they will depend upon the owner's definition of content integrity, as well as their tolerance for risk – both of which may change over time. Costs also depend upon the institutional mission of the owners: is long-term storage necessary, (e.g., for an agreed period of legal retention)? The attributes of original materials will affect the cost of digital preservation: for example, their complexity and quantity.

Other cost factors will depend upon the scope of services required and the preservation obligations of the owners: will there be a need to preserve bits (just the file) only; or to preserve file and associated metadata (e.g., related to intellectual property rights)? Do these need to be tracked and modified over time? Is the intention to preserve use (behaviors, and capability)? Or even to preserve faithful rendering? Deciding which of these aspects need to be preserved for the long term will affect cost.

Chapman stated that the attributes of the materials themselves should not matter in terms of cost, and referenced Kevin Ashley of the University of London Computer Centre, who has declared that digital preservation costs correlate to the range of preservation services that are on offer, not the attributes of the materials. That is, preservation costs will not be the same for identical materials held under different service level agreements at different archives. (See Resources)

Reiterating that the repository is the nucleus of preservation activity, Chapman stated that repositories will be required to ensure the longevity of digital materials. The majority of content owners will become consumers of centralized repository services, therefore repository storage costs — independent of costs for ingest or access — must be affordable or owners will withhold materials from deposit, running the risk of their loss. The price for storage should be what owners can afford to pay. He introduced his case studies: the Harvard Depository (centralized storage in a managed environment), and the

OCLC Digital Archive, a new commercial service, emphasizing that an analysis of the use of existing traditional (analog) repositories is a relevant indicator of what owners will pay for managed storage of digital objects. Chapman presented a "snapshot" cost recovery billing model for each organization, which appears below:

Both use unit costs that have been priced to recover operational costs of actually managing a repository, as a repository provides more than storage. The OAIS model for digital preservation emphasized data management, archival storage and administration. A great deal of infrastructure is involved in managing data, and there are a lot of costs to recover.

Cost-recovery billing model for both: in each case, size is the metric of billing.

Harvard Depository
$3.91 per billable square foot (standard)
$9.85 per billable square foot (film vault)

OCLC Digital Archive ("bit preservation")
$60.00 per GB (1,024 MB), if < 100 GB
$32.00 per GB, if 101-1,000 GB
$15.00 per GB, if > 1,000 GB

As consumers of such resources, we are aware that there will be a real "price" but that the "cost" to the end user is what interests us most.

Explaining the cost gaps between analog and digital, Chapman looked at some additional factors, and illustrated them with a series of slides. The costs examined are the costs of storing high-resolution master copies. The relatively low cost of storing ASCII files suggests that digital storage may become affordable.

Other reasons for the cost gap will include key institutional decisions made by each organization related to their business model, and pricing model. These policies will have an impact on business models, including decisions related to where the materials are actually stored – for example, to retain materials in uncontrolled local storage environments, such as library stacks, or to deposit them to managed repositories. Production choices will also affect business models – for example, preservation microfilming produces two copies that will have to be stored. Storage of uncompressed digital images of book pages can be up to 10 times more expensive then microfilm at current HD and OCLC prices. These costs do not factor in OCLC's volume discount, whereas methods do exist to close the "cost gap" by negotiating and collaborating. Chapman stressed that the most important issue is that there are many variables and contexts that will affect costs. The quality of the files, for example, will greatly affect the

cost of digital storage: uncompressed, 24-bit images will be much more expensive to store. Developers should work to close all cost gaps to make repository storage affordable. The cost gap for audio and video is much higher and therefore more significant.

There are other significant costs not included in the pricing model. Key curatorial decisions also matter, as does the issue of integrity of data – how much material can you afford not to keep? What quality is required, and what format (vis-à-vis use requirements). A level of risk must be assessed (e.g., is compression acceptable?), as will the extent of technical and administrative metadata required.

There are still some open questions related to the issue. The decisions taken by an institution are the key to determining costs. Can institutions afford to digitize at the highest quality technology allows, then keep the digital objects that result from this strategy?

Can institutions afford to keep all versions? Can they afford not to?

Back to top

FROM PROJECTS TO FULL PROGRAMS: INSTITUTIONAL COST ISSUES

Carrie Bickner ***New York Public Library Visual Archives***

See: <http://digital.nypl.org/browse.html>

Carrie Bickner discussed the NYPL's Digital Libraries program, and some of the pay-offs from which the institution has benefited as it has made the transition from individual digital projects to more structured digital programs. Most significantly, the infrastructure that has been developed now supports new projects and initiatives.

She described the process of building a team to support digitization initiatives, emphasizing the broad scope of expertise necessary. The Digital Libraries team has 23 staff, with a metadata team of 5 full-time people and some interns. The NYPL projects require a team well versed in all aspects of digitization and technology infrastructure, including support for systems such as Oracle databases and ColdFusion delivery mechanisms. NYPL has elected to use MRSid software to pan and zoom on high-resolution images, a system that can show a great deal of detail in the images. The digital imaging unit is in the NYPL, and most digitization is done locally. The Library also uses vendors for some initiatives, for example JJT has worked with NYPL both on- and offsite.

Bickner emphasized that the technology development should actively reflect and support library standards, and reiterated that metadata specialists were of key importance to the success of such initiatives. Now that the team is in place, and fully equipped, it is using the infrastructure that has been created to do projects that were not originally within the scope of this initiative. For example, equipment and staff are supporting a number of curatorial projects.

The major project is the NYPL Visual Archive, which was formerly known as ImageGate. The project dealt with over 600,000 images from the four research collections of the NYPL. The collection is comprised of many different types of visual materials, including printed ephemera, maps, postcards and woodprints. The project is moving from the Central Building to the Library for Performing Arts and the Schomburg Center. Bickner noted that it is often easier to move people than the materials. In this case, the project is working with glass-plate negatives of images from the performing arts - photos of actors, actresses, set designs, etc. from the 1920s to the 1960s. As glass plates can't be used in the reading room, this will be the first opportunity for the public to view much of this content.

With database and team in place, the Digital Library team is starting work on other projects (see <http://digital.nypl.org/forthcoming.html>), which will include: American Shores - Maps of the Middle Atlantic Region to 1850 and The African American Migration Experience, which will include both images from the archives and specially commissioned essays on each of 13 phases of the African Diaspora – from the early slave trade to recent Haitian experiences. This project will include materials (some still in copyright) from other repositories (e.g., Associate Press photo archive for the 1980s on the Haitian migration). The Library is still developing rights management methodologies for such materials and a half-time staff position is dedicated to clearing and managing copyright for these materials.

Bickner raised an important aspect that similar projects will have to face in the future. She showed a page of Whitman's own copy of the 1860 *Leaves of Grass*, with his annotations for changes for the next edition. Writers today do the same - but with Microsoft Word, deleting previous versions. How will we electronically display the creative process?

In questions, Bickner clarified that most of this work is done by staff funded by "soft" money, i.e. grant funded positions. This is a concern as the team attempts to develop sustainable digital programs.

Back to top

Tom Moritz ***Toward Sustainability - Margin and Mission in the Natural History Setting***

See Presentation Slides: in Powerpoint (21 MB file); in PDF

Tom Moritz began by observing that natural history museums have unique challenges in creating digital collections. The American Museum of Natural History (AMNH) contains over 34 million natural history specimens and these objects, in turn, may have many associated pieces of information in various formats: how do we devise optimal, strategic solutions for the development of efficient, accessible digital programs, with such a mass and range of materials, especially when faced with the constraints on open access to information, dictated by the market, technology, law and norms (using Lawrence Lessig's terms).

In the now (presumptively) "mature stage" of digital library development, he questioned the common opposition between start-up project and sustainable development. Are we not placing expectations on the digital library that we do not on traditional libraries? We know that academics require robust analog libraries for research, and that institutions are required to sustain them for accreditation – do they have the same expectations of digital materials? Are we trying to do too much with too little – at this time, indirect costs and overhead don't support digital initiatives, and too many programs are funded entirely by "soft" money.

Before continuing with his theme of sustainability, Moritz briefly digressed into further consideration of what James Boyle has labeled "The Second Enclosure Movement," with graphs developed by Lessig in *The Future of Ideas* that show increased use of the term "intellectual property" and rampant growth in the concept of "ownership" of information. The Flexplay DVD, that self-destructs 36 hours after opening, is a graphic example of technology enabling this land grab.

As one response to this threat, Moritz described an open access, “common knowledge” project: “Building the Biodiversity Commons,” about which he had written in Dlib magazine. For more information about this project, see <http://www.dlib.org/dlib/june02/moritz/06moritz.html>.

Returning again to Lessig’s ideas on the constraints on open access (the market, technology, law and norm), Moritz raised the concepts of "Mission" and "Margin" as they relate to an organization like AMNH. While it is clear how all four Lessigian elements apply to commercial enterprises, how do they apply to a nonprofit, driven by a “non-commercial” mission? How does digital fit into a mission that was developed for an analog world? The mission of AMNH, as stated by the New York State Legislature in 1869, is to furnish “popular instruction.” This endorses the notion of freely available information. However, in difficult financial times there is pressure to generate revenue for all organizations and projects, and more objections to open and free access to digital content.

While AMNH explores ideas for mission-consistent revenue generation options, there are presently several potential and actual sources of revenue that are generating funds for different sectors of the organization:

- Licensing images
- Sales of images to the luxury market and photo sales
- Scientific Publications
    - Sales of print (subscriptions/ single issues)
    - Royalties from value added databases (e.g.,BioOne)
- Digital Consultancy
- Salary Relief
- Interlibrary Lending
- Grant support from funders (with its attendant “hamster wheel” of constant, grant-seeking activity).

Moritz moved towards a conclusion by asking what the core of natural history is and what are the discipline-specific objectives that can be supported by the digital library, and that will facilitate scholarship in this field? Strategic developments should be informed by close analysis of the requirements of academic research, and should provide a conceptual framework to provide integrated access to publications, archival records, field notes and specimens. Content should be both widely distributed, and strongly integrated. A 1998 article in Nature suggests that there are 3 billion specimens in 6,500 Natural History museums around the world. This variety of content – not merely specimens and artifacts, but also field notes, images, formal publications, exhibit labels, etc., requires careful cataloguing.

The Darwin Core (DwC) has been developed as a "discipline-based profile describing the minimum set of standards for search and retrieval of natural history collections and observation databases". But such solutions need to be efficient and parsimonious. The Semantic Web makes possible an ontologically-based solution applying formal, explicit specifications of a shared conceptualization of ”natural history”. Projects such as the AMNH digital collections related to the Congo begin to illustrate these ideas. See <http://diglib1.amnh.org> and <http://library.amnh.org/diglib/resources/index.html>.

A question raised a very important issue, inspired by the emphasis on the need for shared developments, natural history registries, protocols, etc. The community requires collaborative initiatives, but instead we are all in competition for private funding. The collaborative potential of our shared skills and talents needs to be addressed at the community level.

Back to top

Steven Puglia ***Revisiting Costs***

See Presentation slides: as pdf

Puglia’s presentation firmly focused on actual costs, and updated some of the data presented in his 1999 article: “The Cost of Digital Imaging” <http://www.rlg.org/preserv/diginews/diginews3-5.html#feature>.

He emphasized that there are many costs involved in digital imaging projects, of which scanning is only a part. Costs will be related to: the selection and preparation of originals; cataloging, description and indexing; preservation and conservation; production of intermediates; digitization; quality control of images & data; network infrastructure; on-going maintenance of images and of data.

Puglia’s examination of overall average costs stemmed from his experience on a number of grant review panels. But to make any decent comparative study he obviously needed access to a range of material – and that has not been easy. He had access to cost information from the National Archive’s Electronic Access Project but there were virtually no published reports on costs from other projects. Neither funders nor project managers have access to useful metrics on the cost of digitization to guide them. And when information is available, comparing it is next to impossible: most cost models are not sufficiently granular and there are lots of hidden costs, especially in the interstices. In order to validate a particular cost model, each step in the conversion process must be

articulated in detail. Puglia emphasized that as costs vary so much, what is most important is their range, and he featured this in his presentation.

Overall, he noted that on average, roughly one third of the costs are related to digital conversion, one third for cataloging and descriptive metadata, and one third for administration, quality control, etc. In his 1999 article he quoted an average cost, over three years of data, of $29.55 per digital image (but with a range of between $1.85 and $96.45). Within that, itemized average costs come to $6.50 for digitizing; $9.25 for cataloging; and $13.40 for administration. Adjusted for unrealistically high or low costs, the figures came to $17.65 overall (digitizing $6.15; cataloging $7; and administration $10.10). [See presentation slide 5].

The Library of Congress National Digital Library Program originally planned to digitize 5 million digital images over 5 years for $60M, which would be $12/image – although at one point NDL had 85 people on staff, which would increase overall costs. On examining the NDL annual report for 2001, we can conclude that the project has actually produced 7.5M images, as some 25% have 2 images or versions, 25% have 3, and 25% have 4. Thus there are about 3M unique items or images in the NDL and the cost is really $20/ image. This cost does not include the Ameritech collections, which are about 20% of the site, and it had $1.75M over three yrs. Partner institutions paid the rest of the cost - so actually the numbers are low because it doesn't include partner costs.

In figures from the NDL annual reports, Puglia showed that of the $43M grant for the NDL over 5 years, 46% went to personnel, 27% for digitizing & services, and 18% on professional and consulting services.

The Library of Congress reported, in the *Report of the Task Force on the Artifact in Library Collections* (CLIR, 2001), that it spent $1,600 per book, or $5.33 per page, for base level digitization. Enhanced digitization was $2,500, or 8.33 per page, but this figure was based on the costs reported in Puglia's article.

Questia Media reportedly spent $125M to digitize 50,000 books, or $2,500 per book. *Forbes Magazine* (April 2, 2001) stated that it would take $80M, or $200-$1,000 to scan and proofread each book. In an article on Questia in the *Chicago Tribune*, the Questia CEO was quoted as saying that it took $100M and 2 years to get 40,000 books online and 20,000 in production, at a cost of $1,700-$2,500 per book (however, a commentator at the end of the session pointed out that Questia's costs must factor in enormous marketing and advertising costs, which would bring down their overall cost per image).

A brief review of other data that Puglia had surfaced included:

- The National Yiddish Book Center: $3.5M for 12, 000 pages, or $292 per book.

- Corbis' Bettman Archive scanned at a cost of about $20 per photo (stopping after 225,000 photos). Of its 65 million images, Corbis has about 2.1M digital images online for licensing.
- Denver Public Library reports $18-20 to digitize & catalog a photo, including preparation, research, cataloguing and scanning, but not selection, curatorial decisions, equipment, or administrative costs. The project at Denver is scanning about 52 photos per day.
- Boulder Public Library is scanning the Carnegie Historical Images collection for $15 per image.
- Stuart Lee, in a recent article on digitization costs, states that a small manuscript with 200 sheets costs $3,000 overall ($1,000 to digitize).
- Virginia Historical Inventory reported $24.45 per item (including $7.20 to catalog one photograph and $8.16 to digitize it, while a map costs $109 to digitize)

Puglia asserted that ongoing costs are key and must be planned for from the beginning. Minimal maintenance of one set of master image files and access files will be 50-100% of the initial investment for the first ten years; larger repositories might be able to drop this to 10-25%. Cost to install, staff and maintain network infrastructure and digital data for 1st ten years is 5 times the initial investment. In the IT world, the full lifecycle cost is 10 times the development cost.

Retrospective digitization initiatives can only justify the maintenance of images that are actually used, and will need a rigorous cost benefit analysis to assess if this is worthwhile. This can be assessed by use – for example, NARA had 6.7M hits per month - 2.3M hits/month on the Exhibit Hall, 1/3 of all hits. 46,000 search sessions per month, 12 searches/session. This is compared to 6,400 onsite researchers, 35,000 oral inquiries and 31,000 written inquiries per month for over 20 facilities nationwide.

Back to top

Jane Sledge ***Challenges in Storing Digital Images***

See Jane Sledge's paper: download in MSWord

Jane Sledge illustrated the importance of developing workflows and methodologies for generating and storing high-definition images. Digital imaging is becoming an integral part of the National Museum of the American Indian's collection management and outreach activities, and is used for about 85% of photographic activities. Staff use digital cameras in their day-to-day work to prepare condition reports, take preliminary conservation images, prepare high definition images for exhibitions and publications, and

document public programming events or generate images as part of public programming activities.

She focused on a specific collections documentation project in support of a key institutional objective: the digital imaging of NMAI's collections as part of a move of these collections from the Research Branch in the Bronx to the Cultural Resources Center (CRC) in Maryland. The project created a visual documentary record of some 800,000 objects managed by 250,000 electronic records. If an object is lost, misplaced, stolen, or broken in transit, NMAI has a documentary image to show the object's condition at the time of packing and prove that it was in the possession of the museum. The images enable staff to plan and organize both exhibit development and interactive exhibits planned for certain areas. Because all objects are digitally photographed as part of the move process, the overall cost of the imaging project - $2.5 million - is much lower than one driven by an "on demand" process. These costs include the costs of storing the images.

For each image, two sets of TIFF files are stored on DVD-RAM. One is stored at the Research Branch in the Bronx in a fire-proof safe, the second is sent to the Photo Services Department at the Cultural Resources Center in Maryland, then loaded to a Storage Area Network (SAN). A low resolution JPEG copy is also made. Staff send the JPEG file over the Smithsonian Institute network with the move system data. They also link to the Registration Information Tracking System (RITS) application.

NMAI faced an image storage challenge not yet tackled by most museums. Technology staff estimated that images generated by the move project would be in the order of 250,000 on 500 DVDs. A single NMAI TIFF image can range between 10 and 20 megabytes (MB) in size and one DVD can store about 10 gigabytes (GB) of images. Technology staff estimated that the TIFF images might require about 5 terabytes of storage space in an on-line environment. The economics of creating the images is one thing, but finding a sustainable economic framework for storing them is another matter. The options for holding the TIFF images in an on-line environment and linking these to electronic collections' records to provide access were evaluated, and on the basis of this, NMAI acquired a relatively new technology known as a Storage Area Network (SAN) to store large volumes of data.

Sledge recounted how a sequence of errors caused a failure of the SAN, resulting in a serious interruption to the project's workflow. Some factors that led to the equipment failure included:

1. NMAI's Network Administrator had training to operate the SAN in normal situations, but had insufficient training to operate the SAN in an emergency and had been given wrong advice on what to do in the event of a problem.
2. Despite a significant annual maintenance contract, NMAI's SAN was one version

behind in its software updates.
3. The system is designed to provide advance warning to the Network Administrator via email when a disk drive failure is threatened so that preventative action can be taken. This did not happen.

Furthermore, Sledge explained that the system failure was compounded by a project workflow incorporating insufficient back up processes for high-definition imaging projects (e.g., re-using backup tapes). This was due to an over-reliance on the manufacturer's claim that the system had built in fail proofs. Another lesson learned was the importance of understanding and reviewing the back-up plans and procedures in detail (NMAI has subsequently revised and upgraded its backup systems). Sledge noted that there is a need for pro-active risk management and planning at the outset of any digitization project, and that staff's ability to deal effectively with problematic situations should be tested regularly.

In order to reconstruct the data, the project looked to their archival DVD's, only to discover that DVD technology had changed since NMAI first began to store images on DVD-RAMS. Ultimately, NMAI developed a "workaround" solution to recreate the lost data. Michael Lesk (The Internet Archive) commented at the end of the panel session, that unlike a fire or the willful destruction of a library or archive, NMAI was fortunate in that it had multiple copies of the images in a diversity of media and could recover from this misfortune.

Based on their growing amount of electronic media use, NMAI will carefully consider on-line and near-line technologies, and consider tape storage for rarely used media. NMAI also recognizes the important of migrating digital media storage on a diversity of media. Photo Services staff work closely with Technology staff to review options and select new DVD technologies to create additional sets of Move TIFF images on DVD. NMAI has incorporated digital media management into its collections management policies and has established policies for the deposition of digital media into its archives.

NMAI has since applied "Integrated Project Team Techniques" to its overall Media Asset Management project and has staffed a project team with a mix of program area and IT personnel to recognize the complementary roles of project sponsors, managers, decision makers, end-users, IT infrastructure system engineers, and supporting organizations. In choosing to maintain and store high-resolution images, NMAI is committed to professional management, on-going staff training, timely equipment renewal and maintenance, and strong back-up procedures.

Back to top

CHARGING THE CONSUMER

Christie Stephenson Expanding Local Programs Through Revenue Generation

See presentation slides: in Powerpoint

Stephenson described the efforts at the University of Michigan Library to support and expand local conversion efforts by supplementing base funding with revenue generating activity. Revenue generation methods range from straightforward fee-based services to some creative multi-institutional funding models for large projects.

Digital Conversion Services (DCS) is one of four units of the Digital Library Production Service (DLPS) in the Digital Library Services Division of the Library. It provides a variety of conversion services, including bitonal scanning, OCR, continuous tone image scanning and photography, and text encoding. Staff size varies according to the volume of work, with additional staff hired if grant funds are available.

DCS's core work is digitization of the Library's own collections. During lulls in its internal workload, DCS services (such as the provision of a full-time photographer/digital imaging technician) are available to other University units and non-profits, on a fee for service basis. These clients can take advantage of the group's expertise and avoid the acquisition of costly equipment, and DCS can leverage its investment in staff, training and equipment. In many cases, DCS will also host content for clients and provide access through the DLPS federated image delivery system.

DCS has continued to grow its program around the assumptions that they can utilize excess capacity during slack times, leverage investments in specialized and expensive hardware and software, and offer the services of highly skilled technicians to their own and other institutions. In addition, they have been able to respond to special opportunities by adding staff tied directly to the revenue potential of those projects.

External clients have included Early Canadiana Online, the Library of Congress, Harvard, Northwestern, the ACLS History E-Book Project and the University of Chicago Press's Bibliovault Project. DLPS is about to embark on a ten year project where it will provide OCR conversion for a projected 100-million page images from the Law Library Microform Consortium, to be put online using Michigan's digital library software, DLXS. The target throughput is over 800,000 pages a month. DLPS also provides some project support for the Early English Books Online Text Creation Partnership or EEBO TCP, a collaboration between ProQuest, the University Libraries of Michigan and Oxford, and the partnership members.

Digital Conversion Services uses a variety of pricing models across these projects. The

fee-based services are firmly grounded in the cost of doing business. For each service, DCS has an established recharge rate, based on a relatively standard formula.

Annual labor costs (salary + benefits) are added to the amortized cost of equipment and specialized hardware and software. This produces an annual cost for the digitization method. They then use an average hourly throughput figure (based on either a sample or actual data) multiplied by 1600 hours (the DCS figure for the number of working hours per year) to establish an annual throughput. By dividing the annual cost by the throughput, they arrive at a per-unit cost for each conversion method. For external customers, 30% overhead is added to the unit cost. Rates are refigured each year and submitted to the University's Office of Financial Analysis for approval.

DCS is also exploring volume-sensitive pricing schemes for larger projects, and other pricing models, such as the partnership structure currently used to fund the EEBO-TCP.

Stevenson listed some of the challenges faced by Michigan in its efforts to explore new funding models for digital conversion, and as they "learn to be a vendor":

1. Michigan's internal conversion methods are highly standardized and tied to its own delivery system. This requires negotiations with clients to ensure that the framework will support what they want to do.
2. In the academic sector, it is difficult to find adequate support for the business processes necessary to support such initiatives. Billing, tracking payments and negotiations with clients are time consuming.

One of the biggest drawbacks to organizing around even partial dependence on revenue is the uncertainty that comes with it, and the insecurity this can create for staff on short-term contracts.

On the other hand, providing conversion services for external customers can be rewarding, and there is a potential for real collaborative opportunities. Stevenson concluded by looking to models such as the UK's Higher Education Digitization Service (HEDS), which has shown that the presence of a community mandate, the provision of adequate business support and the removal of at least some of the uncertainty might result in a more viable model—and certainly a "learning" model where customer and service provider might explore new methods together to achieve a better result. How such service centers might emerge in our decentralized environment and how they would be managed remains an open question.

Back to top

Kate Wittenberg ***Sustainability Models for Online Scholarly Publishing***

For presentation slides: see Powerpoint

Wittenberg focused on the issues involved in creating and sustaining a stable and effective scholarly publication.

First, she introduced some basic questions related to sustainability:

1. How much does it cost to develop the editorial and technical basis for such a publication? This will require a great deal of contributed skill and time.
2. How much will it cost to maintain a stable and effective publication?
3. How soon will funds be needed for the publication to be stable and continue to be available to users on a regular basis? Often, we neglect this issue until it is too late.

Wittenburg then listed four potential sources of revenue, and the implications of each:

1. *Institutional subscriptions* – this can be a good source of revenue, as long as a project is specific about the resource that is being charged for, and who should be charged. No-one will want to be seen charging third graders for electronic content, but it may be acceptable to charge their school, or school district. This model will require the support of marketing, billing and accounting staff.

2. *Individual sales* – in scholarly publication, this is not an easy route. Individual book sales are poor, and will not sustain a resource, so this should be seen a supplementary source of revenue only.

3. *Foundation support* – it is getting harder and harder to attract grant support, and relying on this source leads to what Tom Moritz identified as the “hamster wheel syndrome” – never being able to step off the grant writing treadmill for long enough to do anything else. Grant writing staff will be required to support this model.

4. *Institutional support from the host institution* - whether universities, schools, museums, etc. Projects will be strengthened if they are supported as a core part of the organizational infrastructure. This is becoming difficult given the present financial situation, and such arrangements usually have to be made at the very start of a project. Staff will be required for negotiation, billing, and accounting. Unfortunately, many institutions are rigidly organized and there is little connection between various programs and departments. It is extremely hard to get “interdisciplinary” projects under way,

especially given the complex decision-making processes and necessary buy-in at libraries and universities. The key managers who are empowered to make these decisions often do not talk to one another.

Related issues concern the timing or launch of such resources – when can they be judged to be ready for release, sale, or new funding? What work has to be done before a business model can be developed? What editorial and technical development must be completed? No part of the project should be in an experimental phase when resources are launched. Furthermore, how will any project partners be involved in matters relating to revenues, collaboration, IP protection? Again, this will have to be thought out very early on in the project life cycle.

Decisions that must be made before a product can be launched include

- What kind of staff are needed, and at what phase in the project – marketing people may be required at launch and during the first year. Thereafter different types of staff will be required.
- When and with whom do we discuss needs for creating a revenue stream? How much of the project will require continual support?

There will also be several long-term questions. How will the business models suggested above affect a project's technical or editorial development (for example, are advertisements acceptable within the resource? If so, from whom and where should they appear?) How will success be measured? How can you change the business model if required? Can the project lower the overall costs of doing business, such as by merging with another partner, or by outsourcing some aspects of the business plan to other places or people? The situation is constantly changing, and developing sustainable business models is an important and ongoing activity.

Back to top

THE ROAD AHEAD:
Jack Abuhoff *A Final Word*

As CEO of the organization that co-sponsored the event, Jack Abuhoff offered some observations on the day's presentations, and gave the audience his sense of some of the critically important points that had been made by earlier speakers:

1. Maria Bonn's warning to watch out for "ramping-up" costs, which have the potential to

derail budget predictions. We tend to base pricing estimates on “steady state” models, when what we need are more dynamic models that predict costs accurately. By paying attention to this issue at the start of a project (even if this will delay the actual project starting point) by document analysis and observing business processes, it will be possible to keep costs – and workflows – under control and prevent ramp-up costs.

2. Nancy Harm’s acknowledgement that Luna Imaging had made mistakes, which indicates that we should want to work with Luna or other vendors who acknowledge and learn from mistakes

3. Also from Harm’s presentation was the message that clearly defined project goals are critical. Project managers should not compromise or accelerate early planning processes.

4. As Steve Chapman pointed out, the community must adopt preservation strategies to enable subsequent users to work with digital resources in the same way that they would be able to continue to work with older, analog materials. This begs the question of whether or not we can afford to scan at a low resolution, or to make other compromises in the digitization life-cycle.

5. Another point made by Chapman was the need to guard against obsolescence – the need for “future proofing”. As technology develops, and costs for bandwidth, for example, decrease, we will see an increase in user demands of electronic information. Much of this will be driven by the emergence of new technologies like the semantic web, which will require changes in the structure of information. We will need our repositories to work in this new environment, and should not feel constrained by the limitations of today.

Finally, Abuhoff explored the metaphor of “home heating costs”. In doing an internet search for this term, Abuhoff had come up with several “hits” – from "99% efficient vent-free gas burners” to “oil burners”. To the consumer, the only concern when shopping for home heating is cost - seventy degrees of heat, from whatever source, will be good enough for the consumer. Digitization is not like this. There are qualitative considerations and benefits to the end user. Digital resources are not consumed immediately, so they will have to be future proofed. We should not approach digitization as buying fuel oil, where the cheapest is the most desirable. Vendors can help the client evaluate what they really want, and what quality they can truly afford.

Back to top

Michael Lesk The Future is a Foreign Country

In addressing the question of how to pay for digital libraries, Lesk invoked Voltaire: “the best is the enemy of the good”. Doing some things really well makes then too expensive for many institutions. Lesk observed that in discussing prices, speakers at the symposium had presented a huge variety of prices for digitization. While it may seem reasonable to spend thousands of dollars to digitize an important cultural artifact like the Beowulf manuscript, how much should we expect to pay to digitize the books used for the Making of America project – books which, Lesk pointed out, would be of no interest to many used book stores. There has been little research on what users really need from digital resources, but some work has been done – Lesk cited some research by Michael Ester of Luna Imaging into the image resolution that is acceptable to users, and it is less than one would expect (see Michael Ester, "Image Quality and Viewer Perception," *Leonardo*, vol 23, no. 1, pp 51-63 (1990)).

He then cited the work of the Carnegie Mellon/Internet Archive's Million Book project, established with the mission of digitizing a very large body of content. Scanning is outsourced to India and China, where inexpensive scanning techniques will be used to produce a very low “cost per page”. The goal of this project is quantity, not quality, and this raises the issue of what users really need from digital resources. Lesk referenced the commonly held opinion that, after investing resources into a digitization project, one shouldn’t have to scan again in 5 years. He argued, however, that if there is a demand for higher quality scanning, the demand itself should help facilitate the necessary funding, and newer technology should makeit easier and cheaper (assuming the copyright situation hasn't deteriorated in the interim).

Lesk returned to an earlier theme, introduced by Don Waters, of being able to assess the “benefit” of doing something, as well as its cost. Having built the best analog libraries in the world, how can we now develop the best digital library systems? How will it be possible to make the systems work, and work with each other? And what will be the cost to smaller libraries if large research universities are able to digitize their entire library collection and put them online? Will a smaller institution still need to have a library to become accredited? Will it be worth maintaining small libraries if large research collections are available online in their entirety? And what are the economics of this? We are now able to have services on the desktop that, until very recently, were only obtained by physically going into a library. What is the cost to the library of offering this sort of service online at no charge to the user? And what is the saving to the institution of no longer having to provide other traditional services?

Answers to questions of this nature can be found in addressing the way people work with analog resources, and the benefits of traditional libraries. Overall, we need to understand users and the patterns of use in order to gain the greatest benefits from our future electronic resources.

Ending though with a demonstration of the critical importance of the library, Lesk cited the story of Sir Alexander Fleming and the discovery of penicillin. Fleming (a doctor) first discovered that some substance from the mould *Penicillium* killed bacteria in 1928, and wrote a paper about the substance, hoping for help from a biochemist. But little happened for over a decade. Prompted by the Second World War to look for antibacterial agents, Sir Ernst Chain, a researcher at Oxford, found Fleming's 10-year-old paper in the British Journal of Experimental Pathology. This discovery in the stacks led Chain and Lord Howard Florey to test and then exploit the first modern antibiotic, to the great benefit of medicine and humanity; Chain, Florey, and Fleming shared the 1945 Nobel Prize. Libraries let us accumulate wisdom for later use; this must be preserved in the digital library of the future.

Back to top

BACK