NATIONAL INITIATIVE FOR A NETWORKED CULTURAL HERITAGE

ninch

The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials

by the

Humanities Advanced Technology and Information Institute, University of Glasgow

and the

National Initiative for a Networked Cultural Heritage

http://www.nyu.edu/its/humanities/ninchguide/

For HATII

Seamus Ross
Ian Anderson
Celia Duffy
Maria Economou
Ann Gow
Peter McKinney
Rebecca Sharp

For NINCH, 2002

President: Samuel Sachs II
President-Elect: Charles Henry
Executive Director: David L. Green

NINCH Working Group on Best Practices

Chair: David L. Green
Kathe Albrecht
Morgan Cundiff
LeeEllen Friedland*
Peter Hirtle
Lorna Hughes
Katherine Jones
Mark Kornbluh
Joan Lippincott
Michael Neuman
Richard Rinehart
Thornton Staples
Jennifer Trant**

**** through June 2001***
***** through May 1999***

Version 1.0 of the First Edition, published October 2002
Version 1.1 of the First Edition, published February 2003

The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials

Table of Contents

Preface and Acknowledgements

I am delighted to introduce the First Edition of the *NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials*. Since the *Guide* was first imagined and seriously discussed in 1998, much committed thought, imagination and expertise have gone into the project.

Back then it was clear that high-level guidance was needed (engaging multiple perspectives across different institution types and formats) to make sense of the plethora of materials coming out on information and technical standards, metadata, imaging, project management, digital asset management, sustainability, preservation strategies, and more. NINCH had been created in 1996 to be an advocate and leader across the cultural heritage community in making our material universally accessible via the new digital medium and this project seemed tailor-made for our new coalition.

Following NINCH's own good practice, the NINCH Board organized a working group to consider the best ways to proceed. That group is at the core of this project. We have lost and gained a few members along the way, but they are the *Guide*'s heroes. Let me name them: Kathe Albrecht (American University), Morgan Cundiff (Library of Congress), LeeEllen Friedland (The MITRE Corporation, formerly Library of Congress), Peter Hirtle (Cornell University), Lorna Hughes (New York University), Katherine Jones (Harvard Divinity School), Mark Kornbluh (Michigan State University), Joan Lippincott (Coalition for Networked Information), Michael Neuman (Georgetown University), Richard Rinehart (Berkeley Art Museum/Pacific Film Archives, University of California, Berkeley), Thornton Staples (University of Virginia) and Jennifer Trant (AMICO). Archivists, librarians, scholars and teachers, digitization practitioners, visual resource experts, museum administrators, audio and moving-image engineers, information technologists, pioneers and entrepreneurs: all were represented in this group. Their expertise, good humor, persistence and good judgment have been essential to our producing this material.

After defining the project and declaring our core principles (detailed in the Introduction), the Working Group issued a Request for Proposals to conduct research into the state of current practice and to write the *Guide* in close collaboration with the Working Group. Of the several fine proposals submitted, we selected one from a broad and experienced team from the University of Glasgow. Under the leadership of Seamus Ross, a research team, based at Glasgow's Humanities Advanced Technology and Information Institute (HATII), mapped out an ambitious survey of the field for gathering information about current practice in the selection, planning, digitization, management and preservation of cultural heritage materials. We thank them for their work.

Although the *Guide* is the heart of this resource, the online version (http://www.nyu.edu/its/humanities/ninchguide/) includes a general bibliography compiled by HATII together with the reports on the 36 interviews that formed the chief

armature of the research underlying the *Guide*. I want to thank the 68 practitioners who offered us their experience and wisdom.

With a working draft in hand, the NINCH Working Group invited a team of volunteer, expert readers to consider our product. They probed and critiqued, and added richly to the text. Let me thank Melinda Baumann (University of Virginia Library), Stephen Chapman (Harvard University Library), Barbara Berger Eden (Cornell University Library), Georgia Harper (University of Texas), Sally Hubbard (Getty Research Institute), Leslie Johnston (University of Virginia Library), Amalyah Keshet (Jerusalem Museum, Israel), Deb Lenert, (Getty Research Institute), Kama Lord (Harvard Divinity School), Alan Newman (Art Institute of Chicago), Maria Pallante (Guggenheim Foundation) and Michael Shapiro (U.S. Patent and Trademark Office) for their readings and contributions. All who have read his comments would quickly agree with my singling out Steve Chapman as one who exceeded all of our expectations in the depth of his reading and the comprehensiveness of his responses. So a special thank you to you, Steve: we are indebted to you.

Julia Flanders, of Brown University's Women Writers Project, served as an inspiring copy editor, going far beyond what we might have asked of her.

Lorna Hughes, Assistant Director for Humanities Computing at New York University, arranged for the generous donation of web services to mount this edition of the *Guide to Good Practice* on the Internet. Antje Pfannkuchen and Nicola Monat-Jacobs have done a superb job of tirelessly mounting many pre-publication versions of the text online leading up to this final First Edition: we thank them heartily for their accurate and prompt work. Meg Bellinger, Vice President, OCLC Digital & Preservation Resources, has offered the services of that division in mirroring the *Guide* on OCLC web sites in the U.S. and abroad and in furthering the *Guide*'s further development. Thanks to Robert Harriman, Tom Clareson, Judy Cobb and Amy Lytle in making that happen.

Many thanks to the Getty Grant Program for initially funding this project and making it possible.

For all of its richness and complexity, we propose this as the first of several editions of a living document. Future developments and discoveries will add to and refine it. What can your experience add? The Second Edition will incorporate not only your comments but also an online navigational system based on a set of decision trees that should dramatically improve access to the information and advice.

Please use our Comments Form to update or correct information or suggest features that will enable us to make the Second Edition increasingly useful in assisting this broad community to network cultural resources more effectively:
http://www.ninch.org/programs/practice/comments.html

David Green
October, 2002

I. Introduction

The Case for Good Practice

Early developers of digital resources often had little thought for how their projects might dovetail with others. Today many of these projects suffer from this lack of forethought; they cannot be extended for broader use, they cannot be built upon by others and the chances are slim that they will survive into the future. More recently, the cultural community has begun to realize the importance of applying technical and information standards intelligently and consistently. The use of such standards not only adds longevity and scalability to the project's life cycle, but also enables an ever widening public to discover and use its digital resources.

One of the goals of this *Guide to Good Practice* is to show the critical importance for the community of moving beyond the narrow vision of these early project-based enthusiasts and thinking through what is needed to establish sustainable programs. By adopting community shared good practice, project designers can ensure the broadest use of their materials, today and in the future, by audiences they may not even have imagined and by future applications that will dynamically recombine 'digital objects' into new resources. They can ensure the quality, consistency and reliability of a project's digital resources and make them compatible with resources from other projects and domains, building on the work of others. Such projects can be produced economically and can be maintained and managed into the future with maximum benefit for all. In short, good practice can be measured by any one project's ability to maximize a resource's intended usefulness while minimizing the cost of its subsequent management and use.

By adopting community shared good practice, project designers can ensure the broadest use of their materials, today and in the future, by audiences they may not even have imagined and by future applications that will dynamically recombine "digital objects" into new resources.

Within the cultural and educational communities, there are today many different types of guides to good practice written for particular disciplines, institution types or specific standards. These include the Text Encoding Initiative's Guidelines for Electronic Text Encoding and Interchange, Cornell University Library's *Digital Imaging for Libraries and Archives*, the Digital Library Federation's *Guides to Quality in Visual Resource Imaging*, the Getty Trust's *Introduction to Vocabularies* and *Introduction to Metadata* and the UK's Arts and Humanities Data Service series of discipline-based "Guides to Good Practice." In creating the National Digital Library, the Library of Congress has

been assiduous in providing documentation and discussion of its practices; similarly, the National Archives has published its internal "Guidelines for Digitizing Archival Materials for Electronic Access," and the Colorado Digitization Project has brought together in a web portal a wide-ranging collection of administrative, technical, copyright and funding resources.

Link Box:

Existing Good Practice Guides

Guidelines for Electronic Text Encoding and Interchange (Text Encoding Initiative):
http://www.tei-c.org

Digital Imaging for Libraries and Archives (Cornell University Library):
http://www.library.cornell.edu/preservation/dila.html

Guides to Quality in Visual Resource Imaging (Digital Library Federation):
http://www.rlg.org/visguides/

Introduction to Vocabularies (The Getty Trust):
http://www.getty.edu/research/institute/vocabulary/introvocabs/

Introduction to Metadata (The Getty Trust):
http://www.getty.edu/research/institute/standards/intrometadata/

"Guides to Good Practice" (Arts and Humanities Data Service):
http://ads.ahds.ac.uk/project/goodguides/g2gp.html

"Guidelines for Digitizing Archival Materials for Electronic Access" (National Archives):
http://www.nara.gov/nara/vision/eap/digguide.pdf

Various documentation from the Colorado Digitization Project:
http://coloradodigital.coalliance.org/toolbox.html

The Library of Congress has published many supportive materials; some notable resources include:

"Challenges to Building an Effective Digital Library":
http://memory.loc.gov/ammem/dli2/html/cbedl.html,

"Technical Notes by Type of Material":
http://memory.loc.gov/ammem/dli2/html/document.html

"Background Papers and Technical Information":
http://memory.loc.gov/ammem/ftpfile.html

"Manuscript Digitization Demonstration Project, Final Report":
http://memory.loc.gov/ammem/pictel/

"Lessons Learned: National Digital Library Competition":
http://lcweb2.loc.gov/ammem/award/lessons/lessons.html

"Conservation Implications of Digitization Projects":
http://memory.loc.gov/ammem/techdocs/conservation.html

Put simply, this plethora of information is daunting. Where does one start and how does one evaluate the relevance of any particular text in the growing corpus of material on project planning, digitization, the kinds of metadata that need to be included in any project, and the maintenance and preservation of digital resources?

As we detail below, the *NINCH Guide* has a good claim to being unique in providing a broad platform for reviewing these many individual statements. First, it is a community-wide document, created and directed by a NINCH Working Group culled from practitioners from digitization programs in different types of institutions (museums, libraries, archives, the arts and academic departments) dealing in different disciplines and different media. Second, it is based on a set of broad guiding principles for the creation, capture and management of networked cultural resources. And finally, it is also based on a set of intensive interviews of substantial digitization programs in the U.S. and abroad. The perspective is thus a new one.

By offering universal access to the knowledge this research brings together, the *Guide* should help to level the playing field, enabling newcomers to the field and projects which are smaller, either in terms of budget or scope, to offer resources that are as valid, practical and forward-thinking as projects that are created within information- and resource-rich institutions. It is this sharing of knowledge that truly facilitates the survival and success of digital resources.

History, Principles and Methodology of the *NINCH Guide*

The National Initiative for a Networked Cultural Heritage (NINCH) is a US-based coalition of some 100 organizations and institutions from across the cultural sector: museums, libraries, archives, scholarly societies, arts groups, IT support units and others. It was founded in 1996 to ensure strong and informed leadership from the cultural community in the evolution of the digital environment. Our task and goal, as a leadership and advocacy organization, is to build a framework within which these different elements can effectively collaborate to build a networked cultural heritage.

Realizing from the start the importance of connecting the big picture (the overall vision and goals for a networked cultural heritage) with actual practice within cultural institutions, NINCH board and staff concluded that organizing a comprehensive Guide to Good Practice was an important priority. A NINCH Best Practices Working Group was created in October 1998 to organize a review and evaluation of current practice and to develop a set of principles and guidelines for good practice in the digital representation and management of cultural heritage.

The Group proposed an initial definition of good practice by distilling six core principles from their own experience with a set of evaluative criteria to judge current practice. The Group thus proposed that Good Practice will:

1. ***Optimize interoperability of materials***

   Digitization projects should enable the optimal interoperability between source materials from different repositories or digitization projects

2. ***Enable broadest use***

   Projects should enable multiple and diverse uses of material by multiple and diverse audiences.

3. ***Address the need for the preservation of original materials***

   Projects should incorporate procedures to address the preservation of original materials.

4. ***Indicate strategy for life-cycle management of digital resources***

   Projects should plan for the life-cycle management of digital resources, including the initial assessment of resources, selection of materials and digital rights management; the technical questions of digitizing all formats; and the long-term issues of sustainability, user assessment, digital asset management and preservation.

5. ***Investigate and declare intellectual property rights and ownership***

   Ownership and rights issues need to be investigated before digitization commences and findings should be reported to users.

6. ***Articulate intent and declare methodology***

   All relevant methods, perspectives and assumptions used by project staff should be clarified and made explicit.

With funding from the Getty Grant Program, NINCH issued a request for proposals to conduct a survey and write the *Guide*, in close collaboration with the Working Group. A team organized by the Humanities Advanced Technology and Information Institute (HATII) of The University of Glasgow was hired.

In order to ground the *Guide* in the reality of good practice that has been proven in the field, and to ensure that the personal views of the Working Group did not color the *Guide* too much, the project began with a thorough review of current literature on the subject of good practice that included online and print resources, as well as gray[1] literature. This process was complemented by structured face-to-face and telephone interviews, and selective written exchanges with individuals from the cultural heritage sector.

The key information-gathering tool used for research was the Digitization Data Collection Instrument for Site Visit Interviews developed by HATII. For details on the development and use of this interview instrument see the "Introduction" to the Interview Reports. Interviews at digitization facilities lasted between 90 minutes and 3 hours and were conducted by four researchers on 20 site visits, involving 36 projects and 68 individuals from late 2000 through early 2001.

Sites were selected on a "best fit" basis to a matrix of project types and key themes established by the project team. The sites selected were not a scientific or representative sample, but as a group they broadly reflected the diversity of the community, while each represented one or more of the identified key themes of good practice. The rationale for site selection is further explained in the "Introduction" to the Interview Reports.

In parallel to the site visits, the research team undertook further focused research via literature review, telephone interviews and written correspondence on several broad themes: text encoding, digital preservation, asset management, rights management, and quality assurance. HATII identified another set of relevant digitization sites for inclusion in this stage of research. Theme reports written out of this research filled knowledge gaps that had not been addressed by the site visits and provided a more analytical view of current community good practice in these areas.

How To Use the *Guide*

The *NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials* is a unique contribution to the field. It takes a process-oriented approach to the digitization and management of cultural resources (keeping in mind their long-term life cycle from selection through preservation) and does so from a community-wide perspective. NINCH also intends to put into place a system for regular updates and further editions. The *Guide* takes the reader from the identification of available resources and the selection of material, through the creation of digital content, to its preservation and sustained access. For institutions that have not yet begun digitally representing material from their collections or making their born digital material accessible, the *Guide* will provide a way of coming up to speed in a quickly developing area. It identifies the decisions that need to be made, indicates when they need to be made and draws attention to the implications of the possible choices.

Users of the *Guide* will come from different backgrounds. Perhaps five examples will help you situate yourself among the possible categories of readers.

- If you are an archivist, librarian or museum professional, the *Guide* will help you select materials from your collections, reformat them, and make them visible and accessible to different audiences via the Internet or on portable digital media.
- If you are a funder, the *Guide* will give you an understanding of the activities involved in creating, delivering and sustaining digital content and background,

and will help you to assess whether or not requests for funding are sensible and built on a thorough consideration of the issues.

- If you are an academic or other researcher, the *Guide* should give you sufficient information to design a project, convince collection owners to grant you access to material you need to digitize, and persuade funders to support your project.
- If you are a teacher of digitization in a library school or a faculty of information studies, the *Guide* can help you identify central issues to cover in digitization courses, and can provide your students with an understanding of the issues that they will need to address when they join a cultural heritage institution.
- If you are a vendor or manufacturer of software or hardware, the *Guide* should provide you with an indication of the challenges faced by the cultural community and of the significant levels of investment that the community is making in digital content creation, as well as showing you the tremendous value of the intellectual capital with which they are working.

This is not a recipe book for experts or specialists. It will provide content owners and decision-makers with sufficient guidance to know whether or not they are getting the best advice from their technical staff and whether their colleagues have put in place adequate strategies to guarantee the success of their digitization activities. It does not attempt to provide the final word on every topic, but instead supplies links to resources that we have evaluated and have concluded will offer a good next step.

The Guide *identifies the decisions that need to be made, indicates when they need to be made and draws attention to the implications of choices made.*

Humanities and cultural heritage institutions serve the needs of many different communities - from students and scholars to publishers and the general public. As you begin to develop and plan the use of digitization to make your collections visible and accessible, it is crucial to decide which audiences you aim to reach. This will influence many of your decisions: the items you select for digitization, the technologies you will use, and the mechanisms for delivering the digital materials to users. You may find, for example, that you have a collection that interests children as well as adults, but that each audience will require different delivery interfaces. While you could use the same technologies to reformat the material (and you would only need to do it once), and publish both versions using the same underlying delivery system, you would have to develop two separate interfaces to the same material.

Digitization may even change your sense of audience, by making it possible to offer broader access to rare or inaccessible collections. Institutions often think first about digitizing material that is already popular with the public, but digital technologies now

enable them to offer access to material that could not otherwise be seen or used, thus altering rather than simply reproducing the existing profile of use.

Audiences may be not only the users of the digital collections you produce, but also potential creators of digital surrogates from your collection for research, publication, advertising or enjoyment. Examples might be:

- an academic asking to digitize a collection of papers by a recently deceased contemporary artist as part of a research project
- a publisher proposing to produce a pay-per-view website with images of your collection of sixteenth-century engravings of native Americans
- a folk society requesting permission to include a rare recording of a 20th century storytelling from your collection on a CD they hope to release.

How do you respond to these requests?

- What best practices would you require if you were to agree to any or all of them?
- Would your expectations of each project be different or would you set them the same high standards?
- How would you ensure that, while you allow them each to use the material for their different purposes, you retain control of it in digital form, and that the processes involved in its digitization do not put the analog material at risk?

It is worth remembering that analog holdings constitute intellectual capital, and that as digital surrogates are created, the research, teaching or economic value of the originals should not be depleted. This may affect the material you choose to make accessible, the standard to which you do so, and what types of use and access arrangements you will put in place. Requiring those who work with your collections to follow good practices can minimize risks to the analog sources through their digitization.

So the first questions to ask include:

- Where is the audience for my collections?
- What types of individuals does that audience include?
- Will digitization enable me to meet the needs of existing communities better?
- Will digitization enable me to create new audiences for both the digital surrogates and the analog sources?
- What do I mean by “audience” in the digital world? Am I referring only to those individuals to whom I can deliver digital materials or am I also giving consideration to those who would like to produce digital surrogates for business, personal and research purposes?

[1] Gray literature, sometimes called "ephemeral literature," is unpublished material that can be lost to potential readers because it is not disseminated widely through publication or indexing. Examples of gray literature include: government or NGO research reports, workshop or conference papers, and theses.

II. Project Planning

Introduction

Planning is the first and arguably the most important step in any digitization project. Lest this sound like a platitude, it is worth noting that far too many projects are undertaken without adequate thought to the activities involved, the staff required, or the technical exigencies of the work. The need for good planning may be self-evident, but in practice it is often difficult to anticipate all the areas in which forethought is essential. Good planning for any project—even for managers who have successfully completed previous projects—requires a large number of decisions on questions such as the following:

- What work needs to be done;
- How it will be done (according to which standards, specifications, best practices);
- Who should do the work (and where);
- How long the work will take;
- How much it will cost, both to "resource" the infrastructure and to do the content conversion;
- Where, after having answered all of these questions, one might obtain funding.

This kind of planning is one of the most intellectually challenging of the project tasks, and may well be time-consuming. There may also be pressure to hurry this step, from a desire to show visible progress or in response to institutional pressure. But an investment in this kind of planning will be amply repaid over the life of the project: in the quality of the products, in smooth workflow, in staff morale, and not least in the total project cost. The goal of this section is to sketch out the parts of the planning process and indicate the important decisions—assessing the resources needed to complete the project, the staffing and equipment required, the choice and role of metadata, and the overall project management—and how to go about making them effectively. The checklist below gives a brief inventory of the resources required to undertake a digitization project. Not all projects will require all the resources listed, but this list will show the range of needs you should anticipate.

Technology develops and changes so quickly that decisions like those listed above may seem almost impossible to make with any confidence. Information on the array of standards, specifications, equipment, skills, and techniques not only presents a daunting learning curve, but also a welter of detail that can be very difficult to track. For the project planner, however, it is not these details that really inform good decision-making. It is much less important to know what sampling rate a particular piece of equipment

offers than to understand how sampling works and how it can affect the quality of digital conversion. These underlying principles apply more broadly and change more slowly. Most importantly, though, they represent the level at which good planning takes place; with this knowledge, the planner has the tools to bring together an expert group of staff and consultants and create an effective framework within which they can work. This *Guide* contains detailed, up-to-date information on best practices in a number of technical areas, but the *Guide's* greatest and most enduring value for the project planner is its presentation of the more fundamental issues and how they interrelate.

The *Guide's* introductory section has already addressed the first question on the list above: What work needs to be done? By emphasizing the identification of audience and of your own institutional location and goals, the introduction contextualizes this decision and reminds us to ask "Who needs this work? Who will benefit?" The further ramifications of this question are explored in Section III on selecting materials, which discusses how to assess your collections and set priorities for digitization, and in Section XII on user evaluation, which provides guidance on how to assess the needs of your audience and how this information can shape your digitization strategy. This is also the stage at which you should get the facts and make your decisions concerning rights management, without which you cannot proceed with digitization: you need to establish the intellectual property status of the materials you wish to digitize, and you also need to decide on your own strategy for managing the intellectual property you are about to create. Both of these issues are explored in depth in Section IV. And although the project's final product may seem impossibly remote at this stage, you need to consider how the results will be distributed: not only what technologies you will use, but also how you will control access and ensure that you reach your intended audience. Section X covers these issues in detail.

The question of how the work will be done—the specifications, standards, and procedures you need to establish—has many facets which are addressed at various points in the Guide. Foremost among these is the question of standards: by using standards-based approaches wherever possible, you increase the longevity, portability, and interoperability of your data. You need to be aware of the standards that apply to the kinds of digitization you are undertaking, and these are described in detail in the sections on digitizing text, images, and audio-visual materials. Given the complexity and breadth of most standards, though, you also need to be aware of the best practices that apply to your community. For instance, both documentary historians and linguistic researchers use the XML-based Text Encoding Initiative Guidelines to encode textual data, but each group uses the standard in different ways that serve their particular needs. While you are considering the specifications for your data, you should also think carefully about how to capture and represent the metadata you will need to administer your digital materials and enable them to be used effectively. The Guide includes an appendix on metadata which describes the various types and their uses. The relevant sections of the Guide also provide pointers to specific information on best practices for particular digitization communities. The question of "how" also involves decisions about equipment. For the project planner, these questions are most usefully addressed not at the level of specific brands and models, but by thinking about the functionality you require and the tradeoffs you are

willing to make (for instance, whether keeping costs low is more important to the project's overall success than achieving the highest possible capture standard). The sections on images and audio-visual materials discuss how to approach these decisions; more specific information on particular kinds of equipment can be found in the appendix on equipment. Finally, you need to establish an effective workflow for your project. At the highest level, this includes project management strategies, which are discussed later in this section, and quality assurance methods (discussed in Section VIII). But in addition you need to consider how you will store, manage, and track your digital objects, which is addressed in detail in Section XIII on digital asset management.

Staffing issues—who should do the work—are closely related to the points just mentioned, since your decisions about methods and procedures may be difficult to separate from the staff resources you actually have available. Few projects have the luxury of hiring all new staff to an arbitrary standard of skill and experience. Further on in this section we discuss human resources: how to construct job descriptions and identify skilled staff, and how to set up a management and advisory framework that allows your staff the autonomy to do their jobs effectively. In Section IX, Working With Others, we consider a range of collaborative and cooperative relationships that may expand your staffing options, including project consultants, vendor outsourcing, collaboration with other institutions, and internal cooperation.

Once you have worked through the issues sketched above, you will be in a position to assess the practical scope of the project: how long the work will take, and how much it will cost. Of all the questions addressed here, these may be the most vulnerable to change over time, as techniques and equipment improve and grow cheaper, and as quality expectations rise. Some guidance on cost estimation is offered later in this section, and also in the sections on specific digitization areas (Sections V, VI, and VII). You should make sure in researching costs to take into account all of the startup and infrastructural costs the project will incur-costs for initial planning, choosing data specifications, building or choosing tracking and documentation systems, training staff, and so forth-as well as the incremental cost of digitizing the materials themselves. This is also an opportunity to consider the scope of your investment and whether this infrastructure can be reused or adapted for further digitization projects once this project is completed.

Finding the funds to undertake the project is the final step, at least logically; a successful funding request will almost always require a thorough consideration of the issues just described. Even if you are fortunate enough to have funding already committed, going through this process will ensure that you spend your resources prudently and receive value for your investment. Funding sources and strategies are discussed later in this section, and also in Section XI on sustainability.

The checklist box below gives a condensed list of the resources you may need to undertake a digitization project. Although not all projects will need all of the resources listed, it gives a sense of the range and options.

Checklist Box:

Resources that you will need for a digitization project:

Personnel:	advisors project management staff rights specialists researchers editors authors digitizers catalogers technical support/development legal advisors	☐
Software:	operating systems applications: --> image manipulation --> metadata authoring --> database --> indexing and search engine --> web server utilities server systems network clients specialist applications/developments	☐
Storage devices:	local hard drives network storage servers optical devices (e.g. CD writers) magnetic devices (e.g. tape drives) controlled storage environment	☐
Network infrastructure:	cables routers switches network cards ports	☐
Consumables:	stationery utilities printer cartridges lamps (for capture devices/special lighting) storage and backup media	☐
Project management:	preparing bids recruitment publicity and dissemination creation of deliverable product specifications design of workflow supervision of staff quality assurance	☐

Resources within your institution

If you are working within an institution that has other digitization projects under way, an examination of the resources already available within your institution is a good starting point. Staff will know if their department or unit has capture devices available or workers with experience of digitization or cataloging. This is an easy first step towards building a resource inventory, although knowing that you have one flatbed scanner, a digital camera and suitable equipment for digitizing audio, as well as people who know how to use that equipment, is not on its own sufficient. A thorough identification of internal resources involves checking that:

- equipment and software are of a sufficient specification to meet your requirements;
- workers who can operate the equipment are available and appropriately trained;
- technical support and maintenance are in place;
- capture devices are (or can be) directly connected to your storage area; and,
- access to equipment and staff suits your workflow requirements.

Clearly assessing the adequacy of these resources is predicated on other decisions, such as your workflow requirements; indeed, many of the planning areas discussed in this section are closely interdependent. It should also be apparent why the Guide's introductory section stressed early on that you need to define what you want to do and the audience or audiences you intend to reach (see Section I). A clear statement of objectives (preferably in a formal document that can be shared with staff), combined with the resource inventory, will enable you to assess the suitability of your local resources.

You will make this document an even more effective planning tool by adding information about equipment specification (e.g. computer processor speed, RAM, hard disk capacity) and the results of tests for suitability. Before you can conclude that you have suitable resources you must test them to make certain that they will meet the requirements of the project. The Example Box below, "Resource Inventory and Test", shows what a resource inventory and test for scanners might look like.

Example Box:

Resource Inventory and Test:

PCs and Scanners	Functional Requirements	Suitability Test Result
1 Pentium 3, 600 Mhz, 128 MB Ram	Must handle processing and manipulation of image files up to 50 MB	Needs more RAM
1 Pentium 4, 1 Ghz, 384 MB Ram		***Okay***
1 Agfa Arcus		***Okay***
1 Agfa DuoScan 1200		Transparency tray inadequate

Overall Conclusion: Upgrade one PC and replace one scanner

Most large institutions in the cultural heritage sector will have resources that may be useful to the project but would not necessarily need to be borrowed for the entire life of the project. There may be physical equipment, such as devices for digital capture, analog capture equipment (e.g. record, tape, CD and video players that can be used when converting from analog to digital), network storage devices, or handling equipment and controlled storage for analog material.

Human resources may be even more useful—expertise in digitization, text encoding, networks or web delivery can often be found in-house. Even those institutions yet to carry out any significant digitization will have cognate areas of expertise. These skilled individuals can be difficult to find, so tell your colleagues that you are planning a digitization project and have them consider which skills might be of value to you. For example, the skills, techniques and processes required by digital photography are identical in many areas to analog photography, and the same applies to image processing. Similarly, the standards and methods for creating metadata have their roots in the creation of bibliographic records, library catalogs or finding aids and museum collection management systems. In addition to this, it is important to consider the project team and project management process here. Projects should establish a set of procedures for project management from the very start of any project, identifying goals and time scales as well as tasks and outcomes tied to the availability of specific staff and equipment.

It is much easier to identify potential facilities and expertise within the framework of an institutional digitization policy or corporate technology plan—follow the more detailed questions for your own resources as described above. If such a policy has not already been adopted, it will probably be beyond the scope of an individual project to initiate one. Nevertheless, informal inquiries can still be made relatively easily. Remember that apparently unrelated departments or projects may be useful. For example, a great deal of high-end digital imaging takes place in dental, medical, biological and life science departments. The Internal Resource Identification Question Box illustrates some of the common areas of expertise to be found within an institution.

- Is there an institutional digitization policy to adhere to?
- Who else in the institution has digitization projects underway?
- What experience can you use (e.g., photographic, equipment analysis, etc.)?

Question Box:

Internal Resource Identification:

	Institution Type		
Resource	Academic	Library	Museum/Gallery
Imaging	Medical Imaging / Media Services / Photographic Services / Library	Special Collections / Photographic Dept	Imaging / Publications Dept
Metadata	Library	Cataloging Finding Aids	Collection Management Finding Aids
Text Encoding	Literature / Language / Computing Science Depts. / Information Management / Library	Cataloging / Information Management Finding Aids Electronic Texts	Finding Aids / Information Management

External Resources

Identifying resources outside your immediate department, unit or institution can be a more difficult process. Success depends upon what type of institution you are, your strengths and limitations, the accessibility of the resources you are seeking, and whether there is scope for collaboration. Guidance from and access to the experience of others are likely to be readily available. The Link Box points you to national organizations that provide information to support digitization projects. Outsourcing can be another way to fill gaps in the resources available locally, by contracting with a vendor, hiring a consultant, or establishing a cooperative relationship with another institution. These options are discussed in greater detail in Section IX, Working with Others.

Link Box:

Links to National Organizations Offering Guidance

CLIR: Council on Library and Information Resources: "The projects and activities of CLIR are aimed at ensuring that information resources needed by scholars, students, and the general public are available for future generations." http://www.clir.org/

DLIB Forum: "The D-Lib Forum supports the community of researchers and developers working to create and apply the technologies leading to the global digital library." http://www.dlib.org/

LOC: Library of Congress: "The Library's mission is to make its resources available and useful to the Congress and the American people and to sustain and preserve a universal collection of knowledge and creativity for future generations." http://www.loc.gov/

NINCH: National Initiative for a Network Cultural Heritage: "A coalition of arts, humanities and social science organizations created to assure leadership from the cultural community in the evolution of the digital environment." http://www.ninch.org/

RLG: Research Libraries Group: "The Research Libraries Group, Inc., is a not-for-profit membership corporation of universities, archives, historical societies, museums, and other institutions devoted to improving access to information that supports research and learning." http://www.rlg.org/rlg.html

PADI: "The National Library of Australia's Preserving Access to Digital Information initiative aims to provide mechanisms that will help to ensure that information in digital form is managed with appropriate consideration for preservation and future access." http://www.nla.gov.au/padi/

AHDS: Arts and Humanities Data Service: "Create and preserve digital collections in all areas of the arts and humanities." http://ahds.ac.uk/

HEDS: Higher Education Digitization Service: "The Service provides advice, consultancy and a complete production service for digitization and digital library development." http://heds.herts.ac.uk/

TASI: Technical Advisory Service for Images: "Advise and support the academic community on the digital creation, storage and delivery of image-related information." http://www.tasi.ac.uk/

Resource challenges

There are a number of challenges both in assessing and securing the resources required for the project. Projects that take place in large institutions frequently benefit from a significant amount of non-project-related investment. Such hidden benefits include local area networks, high bandwidth Internet connections, large capacity network-based storage devices, web servers, and technical expertise associated with maintaining and developing these facilities. This infrastructure provides the framework for the specific resources and skills a project needs, and without it many projects simply would never get off the ground. Although institutions are now trying to quantify this input, its actual value is difficult to establish, with the result that projects in well-resourced institutions are able to scale up more quickly but often under-represent the real costs that lie behind the their activities.

Equally, less well-resourced institutions and initiatives face an increasing challenge in matching the developments in presentation and delivery of digital resources that larger projects can provide. Frequently, the solution is for small and medium size institutions to develop collaborative projects. The Colorado Digitization Project (http://coloradodigital.coalliance.org/) provides a flagship example of how equipment, staff and expertise can be shared between large and small projects alike, enabling the digitization and delivery of resources that would not otherwise be possible.

Another challenge for digitization projects, large and small, lies in the area of human resources. Content creation is a burgeoning field and although many Internet businesses have failed, those companies such as Getty Images, Corbis, The Wall Street Journal and Reed Elsevier, which have adopted prudent content creation and marketing strategies, are showing steady growth. The finance, commerce, media and entertainment industries all recognize the value and benefits of digital assets, and this places a premium on skilled personnel. Furthermore, the development of staff with digitization skills related specifically to the humanities and cultural field has not kept pace with the growth in the number of digitization projects. Many projects report difficulties in recruiting and retaining staff. Few public sector projects can match the remuneration levels offered by the private sector, but there are strategies you can adopt that enhance your chances of meeting the human resources challenge. These are outlined in the Human Resources Question Box.

Question Box:

Human Resources:

- Are there non-monetary factors that can be emphasized or enhanced? For example, will the project offer advantageous working conditions, training opportunities, or the possibility of gaining qualifications or accreditations?
- Are there aspects of the job that are more attractive than private sector equivalents (e.g. greater creativity, responsibility, freedom)?
- Can posts be combined or split to make most effective use of existing skills?
- Can you consider applicants from a non-humanities/cultural background, particularly for technical posts?
- Can any staff be re-deployed, temporarily transferred or re-trained from elsewhere in your institution?
- Can posts be shared or joint funded with other projects?
- Are you able to outsource any jobs?

Funding

Some project staff will be preoccupied with securing adequate financial resources to start, develop and sustain a project throughout its lifecycle. An accurate picture of the financial costs will help you to identify the financial pressure points and to estimate more accurately the overall costs of running the project. The sections below on skills, equipment, and project management will provide points to help you develop accurate project budgets. An accurate profile of project costs helps to minimize the financial unpredictability of the project and improves the probability that it will attract funding. Funding agencies remain attracted by the opportunities for funding initiatives in the heritage sector. The Link Box provides pointers to some major US funders.

Link Box:

Potential Funders of Digitization Projects:

- Andrew Mellon Foundation: The purpose of the Foundation is to "aid and promote such religious, charitable, scientific, literary, and educational purposes as may be in the furtherance of the public welfare or tend to promote the well-doing or well-being of mankind." http://www.mellon.org/awmf.html
- NEH: National Endowment for the Humanities, "an independent grant-making agency of the United States government dedicated to supporting research, education, and public programs in the humanities." http://www.neh.gov/
- The Getty: "The Getty Grant Program provides support to institutions and individuals throughout the world for projects that promote the understanding of art and its history and the conservation of cultural heritage." http://www.getty.edu/grants/
- IMLS: Institute of Museum and Library Services, "an independent federal agency that fosters leadership, innovation, and a lifetime of learning." http://www.imls.gov/grants/index.htm
- NHPRC: National Historical Publications and Records Commission, "supports a wide range of activities to preserve, publish, and encourage the use of documentary sources relating to the history of the United States." http://www.nara.gov/nhprc/

From the projects surveyed it is evident that most potential funders, particularly in the public sector, require applicants to provide a robust and auditable cost model. How this should be presented may vary from one funder to another, but it can be extremely useful to break down equipment and salary costs on a per unit or work package basis. Not only does it help the potential funders to make comparisons of unit costs between projects within and across heritage sectors, but it also forces you to look at the process and scheduling of work in detail. The accuracy of these figures will be greatly improved by conducting a pilot study or by adopting a cost model from a previous project, even if it needs to be revised in light of the experience of the earlier project.

All the projects surveyed obtained their financial backing from a combination of institutional budgets, public grants, private donation or corporate sponsorship. None of the projects reported serious under-funding, although some found that the distribution of funds created an uneven cash flow, resulting in medium term planning problems. Similarly, none of the projects reported serious concerns about sustainability, even where the source of future funds was unclear. The general absence of plans for self-generating funds or of exit strategies supports this confident view that income would continue to materialize in the future. A number of projects have recognized that failing to adopt long-term financial planning is less than prudent. We recommend that time and support for securing further external funds are crucial as well as exploring the potential for self-generating income. Projects should develop an exit strategy that will secure the maintenance and accessibility of the digital material. These issues are discussed in more detail in Section XI on Sustainability.

Cost models

Determining the cost of digital content creation on a per unit basis is extremely problematic. Not only are there no comprehensive cost models available that cover all resource types but trying to apply such a model to the variety of institution types, financial arrangements, prevailing market conditions, nature and volume of material and the resolutions required would be problematic. Furthermore, the cost basis for creating, storing and delivering digital resources can be quite different and trying to establish a single cost per unit can disguise these differences or ignore them altogether. In spite of these problems it is possible to establish some bases for per unit cost.

At the simplest level a project can take the total funding required and divide it by the total number of units that they intend to digitize. For example total project funding of $300,000 divided by 40,000 units equals $7.5 per unit. However, such a figure can be extremely misleading. Firstly, there will be variation in per unit cost according to the type of material digitized. The creation of OCR text pages will differ from reflective color still images, which will be different again from 16mm moving images or 78rpm records. Even within material of the same broad type there will be variation. Black-and-white negatives are likely to be more expensive to scan than black-and-white prints, since tone reproduction needs to be set image-by-image in the former case, while the same settings can be applied to a group of photographic prints. Even if a project is dealing with material of a uniform medium and size, variations can occur that impact on unit costs. A

collection of bound, legal-size books may have volumes that cannot be opened beyond a certain degree for conservation reasons. This may require a different capture technique, for example capturing pages from above rather than inverted. Some volumes may have details that demand a higher capture resolution than the rest of the collection, while others may require curatorial intervention to prepare them for digitization. The extent to which projects need to take account of such details will vary but at the very least different material types should be distinguished as well as same-type materials that require different capture techniques.

The cost items that go to make up a per unit calculation also require consideration. Should pre-digitization conservation work, handling time, programmers and management staff be included in addition to capture equipment and staff? In practice, projects need to do both. This is best achieved by calculating the costs directly related to capture on a per unit basis, which facilitates comparison and cost effectiveness for different techniques. Non-capture-related items could then be added to provide a total project cost and a second per unit calculation could be carried out if required. The list box below provides an indication of how these different factors can be differentiated. It is common practice to calculate costs for audio-visual material on a per minute basis.

List Box:

Capture Cost Factors:
(per unit for a single media type with uniform capture techniques and settings). It is important to note that the digitization capture costs are actually the least costly of the whole process.

- Handling time (from the shelf to point of capture and return) as a percentage of total salary costs on a daily basis
- Pre-digitization conservation work (this should only be applied for those items that have required it)
- Capture time (from set-up to naming and saving) provided as a percentage of the capture operato total salary costs on a daily basis
- Cataloging/Metadata (required for digitization and/or created at capture stage) as a percentage of total salary costs
- Hardware cost per item
- Quality Assurance time as a percentage of salary cost
- Software cost per item (both hardware and software costs should be on the basis of the depreciation of equipment or projected replacement cost, rather than the total cost of hardware and software)
- Hardware maintenance
- Technical support time (proportion of total salary or contract cost related to capture)
- Project Management time (proportion of total salary related to capture)
- Training (directly related to capture)

Non-Capture Cost Factors*:*

- Additional Project Management salary
- Web Programmer's salary
- Educational Officer's salary (or other additional project staff)
- Cataloging/Metadata (post capture creation) % of total salary costs
- Additional technical support salary
- Additional hardware and software costs
- Consumables (including any storage media such as CDs or DATs)
- Travel and per diem
- Training (non-capture related)
- Storage costs (based on total maintained cost for the gigabytes required)

Some sites with detailed information on costing are listed below.

Key Sites with resources on costings:

- Research Libraries Group: http://www.rlg.org
- Library of Congress: http://www.loc.gov
- Online Computer Library Center: http://www.oclc.org/home/

Human Resources

A project's long-term success depends on the accurate assessment of the required human resources, and producing a map of available and unavailable skills is a valuable starting point. Institutions vary in their areas of expertise and different types of project require different skills. Nevertheless, from the projects that we surveyed it has proved possible to develop a basic template of the people and skills required in realizing a digitization project. The requirements can be scaled according to the size of the project envisaged.

Job descriptions, performance indicators, training

Comprehensive job descriptions are indispensable, regardless of the project or institution. While job descriptions are not always required by the host institution, employment law often demands them. Funders are increasingly expressing an interest in viewing job descriptions as part of the application process as this provides them with a richer overview of the project. It is worthwhile developing an outline of job descriptions before the project reaches the recruitment stage. This is useful to determine the delegation of work, how jobs interrelate, which posts can be tailored to existing skills and which can be identified for external recruitment or outsourcing. A useful process for developing accurate job descriptions is to set out a list of all the tasks required for a post and then rank them from highest to lowest priority or into essential, desirable and non-essential categories. Next, compile a corresponding list linking these tasks to the skills required, including any particular knowledge or qualification. Alongside this, compose a description of the experience or background required for these skills. Finally, review the original tasks and their priority to ensure that a realistic and coherent job description is produced. A resource which has been developed by the Association for Computers and the Humanities is a database of jobs in this field—it may be consulted by projects for guidance in drafting job descriptions, and can also be used to publicize new jobs to a focused audience of candidates. See http://www.ach.org/jobs/ for more information.

Example Box:

Sample Job Description

Job title: Digital Library Research Assistant

The Digital Library Research Assistant will play an integral role in the university's digital library projects, the goal of which is to bring a wide range of source materials to as large an audience as possible. The DLRA has responsibility for overseeing initial scanning and data capture, creating and reviewing metadata, and performing quality assurance checks. With other project members, collaborates on project publications and research.

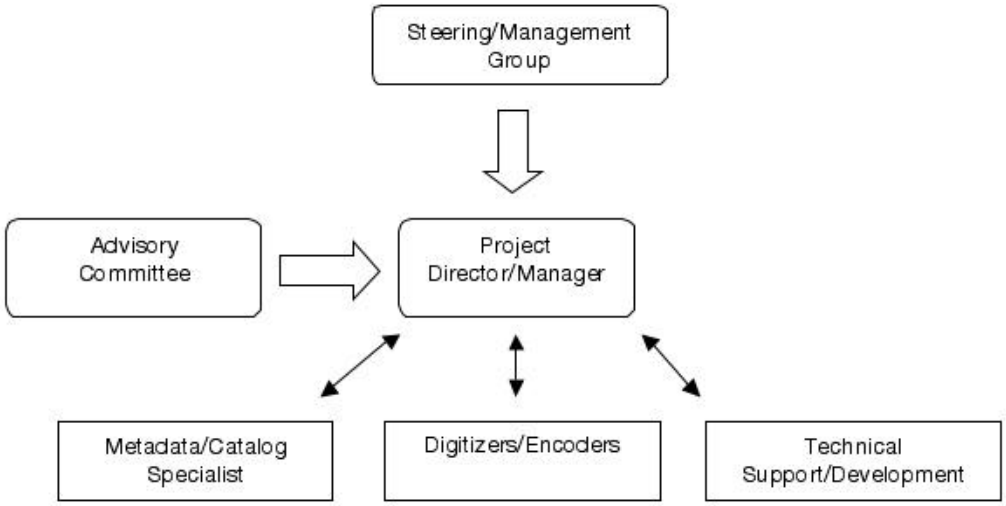
Job requirements: Bachelor's degree and one to three years' experience; basic computational skills, and expertise in at least one area of the humanities. Advanced degree and three to five years experience preferred. Familiarity with relevant encoding and metadata standards, including SGML/XML, METS and Dublin Core, is highly desirable. Must be a self-directed team worker with strong motivation and the ability to take initiative. Needs good communications skills (oral and written) and willingness to work collaboratively.

The use of performance indicators appears to be on the increase. They can have a positive impact, not least by providing a way of formally identifying training requirements. While most projects assess training needs on the job as an informal exercise, formal methods encourage appropriate training solutions to be planned and resourced in advance.

There is a close interplay between performance indicators, job descriptions and training assessments. The job description is very useful in developing meaningful performance indicators. Indeed, a useful starting point for performance review is to evaluate current tasks against those set out in the job description, highlighting whether the original job description was unrealistic, whether workloads need to be re-evaluated in the light of practical experience, or whether a skills shortfall needs to be addressed. The aim of addressing training requirements is to ensure that future tasks can be achieved and that the project will not encounter a skill shortage.

Managing the skills base

It is vital to ensure that a project be able draw on the right balance of skills. The challenge is to determine the skills of individuals and how they can most effectively contribute to the project. The key to successful delivery of projects is management. The diagram below incorporates elements from all of the projects surveyed, from the smallest to the largest, and illustrates the general structure that may be used to manage the project's skills base.



The steering group functions as an executive board and includes all constituents who are directly involved in the project, even if not employed by it, such as curators, archivists, subject specialists and education officers. In practice it is common for the steering group to be an existing committee within an institution.

The advisory committee is a broader-based group, providing general advice on the project's focus and direction. Members usually include the steering group with additional appointments from external organizations bringing particular areas of expertise, such as evaluation, to the initiative. There may be more than one advisory committee, or the advisory committee may be broken down into sub-committees each of which supplies more focused technical, academic or editorial decision-making support. This is the case with the Perseus Project at Tufts University, which has separate Technical and Academic Advisory Boards as well as a Steering Group to provide general project management. (Read Interview 28.2 for details on this arrangement)

It is essential to have a single project manager who is employed by the project, with responsibility for its daily management. In most cases the project manager provides the necessary project management experience, supplemented by internal or external advice. An institution needs to assign both accountability and authority to the project manager position, so that the process is not bogged down by myriad interactions with the advisory

group or groups to deal with daily operations. In content creation projects it is unusual to employ external consultants to handle project management.

What skills are required?

There are four main areas, which will require staff with identifiable skills. These skill areas may be provided within a single project, dispersed across a collaborative project, or outsourced.

- Conservation: A crucial aspect of any digitization initiative will be a conservation assessment of the analog materials. Under some conditions this may show that before some material can be digitized it will require conservation intervention.
- Digitization/Encoding: This can involve digital imaging, keyboarding, OCR, character or full-text encoding, or a combination of these. In some projects it may also include conservation intervention in the analog material.
- Metadata/Cataloging: The creation of metadata records for the digital material. This work may also involve cataloging the analog material or searching for information to enhance the metadata record where it is absent from the analog version.
- Technical Development/Support: This falls into two distinct areas: the creation or implementation of specific IT solutions for creating, managing or delivering the digital material, and the provision of IT support for project hardware and software. This latter area includes desktop applications, network services, and capture devices.

In smaller projects staff may carry out tasks in more than one area: for example, the digitizer may also undertake technical development, or the project manager may take on metadata creation. In larger projects, such as SHOAH or the Genealogical Society of Utah, the duties of staff are so extensive that this is not feasible.

Project managers will have to decide whether to hire new staff with the required skills or to re-deploy existing staff from other areas of the institution. We found that many projects prefer the former, with two notable exceptions. First, there is a discernable trend for photographers to be employed for high-end digitization work. Projects have found that better-quality images are produced through training a photographer in digitization rather than trying to equip a digitizer with photographic skills. The second exception is the tendency to re-deploy or train existing cataloging staff in metadata creation. This is a logical progression for staff who will already have considerable experience in creating bibliographic records, collection management records, finding aids or catalogs, frequently in an electronic form such as MARC.

Another decision concerns background skills. With the exception of some technical posts, we noted a clear preference for staff with arts, humanities, library, museum or gallery backgrounds, or at least some experience or interest in the subject area of the collection.

There may sometimes be advantages in not having such a specialization. For keyed-in text transcription, staff without subject knowledge are more likely to enter exactly what is on the page rather than interpret the contents and enter what they think is in text. On the other hand, subject knowledge can be exceptionally useful in gauging what areas of the content should be focused upon, deciphering difficult materials, or recognizing how areas of the content should be marked up.

When you are trying to find staff with appropriate skills, remember that some projects have benefited from using student labor and volunteers. The ability to draw on student labor represents a significant benefit for university-based projects. Projects such as those based at the University of Virginia Library have been able to build large and diverse digital collections because they are able to draw upon a pool of skilled, motivated and affordable labor. Projects that recruit student labor have invested considerably in training, adopted flexible working practices and tailored the work around the students' educational commitments. This approach has the added benefit of equipping students with the skill set required for future work, adding to the pool of available staff.

Volunteers often provide a similar pool of skills and projects such as the Genealogical Society of Utah have made effective use of this resource. They have found it both necessary and beneficial to invest in appropriate training for the volunteers. Such training should be factored into the project resource plans. In large-scale initiatives, volunteer management and training may become a significant part of the project itself.

The Link Box below provides links to sites that support skills development in digital representation.

Link Box:

An increasing number of organizations are offering training in digitization, which generally proves cheaper and far more useful than commercial training courses:

- Archive Builders: Document Imaging and Document Management. http://www.ArchiveBuilders.com
- Cornell University Library, Department of Preservation and Conservation: Moving Theory into Practice: Digital Imaging for Libraries and Archives. "This workshop is intended for librarians, archivists, curators, administrators, technologists, and others who are contemplating or implementing digital imaging programs." http://www.library.cornell.edu/preservation/workshop/
- HATII: Humanities Advanced Technology and Information Institute. Digitization Summer School: "The course will examine the advantages of developing digital collections of heritage materials, as well as investigate issues involved in creating, curating, and managing access to such collections." http://www.hatii.arts.gla.ac.uk/
- Humanities Computing Unit in Oxford: Summer Seminars covering a range of topics: http://www.hcu.ox.ac.uk/
- School for Scanning - North East Document Conservation Center: Creating, Managing and Preserving Digital Assets. http://www.nedcc.org
- Harvard Extension School, Museum Studies Certification Program. Muse E130: Information Technology for Museums. http: //www.extension.harvard.edu/2001-02/courses/muse.shtml
- TASI: Technical Advisory Service on Images. Training courses "aimed at those involved in digitization projects, those who wish to capture images and those who wish to use digital images in teaching and research." http://www.tasi.ac.uk/training/training1.html
- UCLA/Getty course: Museum, Libraries and Archives. Summer Institute for Knowledge Sharing. http: //skipper.gseis.ucla.edu/orgs/gettysi/html/summer.html
- University of Virginia Library: Digital Media Lab Services tutorials and short courses on digital image, video, and audio capture and editing. http://www.lib.virginia.edu/clemons/RMC/digilab-services.html

Equipment

Because our digitization capabilities are so strongly tied to—and limited by—the developing equipment technology, it is tempting to feel that the available technology should motivate our digitization strategies. However, on the contrary, it is vital to base equipment requirements on the characteristics of the collection/s to be digitized and on project needs, and not the other way around.

Although there are significant cost savings associated with outsourcing work to "offshore" production bureaus in Asia, the Far East, Mexico, etc, in cases where unique materials or special collections materials are to be digitized it is important that digitization should take place as close to the original as possible. Hence many projects will need to confront the complex questions of equipment specification and selection. A

detailed discussion of matching material properties to hardware and capture settings can be found in Section VII on audio-visual materials. There is also further information on equipment choices in the appendix on equipment. At the moment we will focus on the basic differences in equipment and the technologies employed in order that the correct type of equipment resource can be procured for a project. Selecting the most appropriate equipment can be time consuming, but projects should not be deterred by the plethora of manufacturers and their competing claims. For example, the SCAN project (Scottish Archive Network) was initially unable to find a commercially available digital camera that exactly matched their requirements. Instead, they sourced a camera custom-made to their exact specification. This level of exactitude may be out of reach—and unnecessary—for most projects, but it is worth remembering that one need not be entirely constrained by what is commercially available.

Principles of digital data capture

Although there is a variety of capture devices for different applications, whether you are digitizing images, text, audio, video or 3D objects, the operating principles are the same. All digital capture devices take a sample of the analog source material to create a digital surrogate. This sample is made up of two elements: the sample rate and the sample depth. The sample rate describes how frequently readings are taken of the analog material. For example, in a digital image this would be the resolution, or the frequency per unit of area: the number of pixels per inch, expressed as pixels per inch (ppi) or dots per inch (dpi). An image captured at 600 ppi would have had 360,000 samples recorded per square inch. Similarly, for audio-visual materials the sample rate is the frequency per unit of time at which the source material is sampled. The sample depth is the amount of information recorded at each sampling point. For example, a sample depth of 24-bits would capture 8 bits for each of the three color channels (red, green and blue) at every sample point. For a more detailed explanation of sampling, see the appendix on digital data capture and Section VII on Audio-Visual Materials.

Selecting equipment

The medium, format, size, and fragility of the original material are among the primary factors affecting equipment choice. For text documents, flatbed scanners are suitable for single leaf, regular sized documents, provided the material does not go beyond the scanner's maximum imaging area (usually up to approximately US Letter size), or is put at risk by "sandwiching" it in the scanner. Large format flatbed scanners and sheet-feed scanners can handle single leaf, oversized documents. However, sheet-feed scanners put material at greater risk than flatbed scanners as the originals are pulled through a set of rollers. Drum scanners, whose imaging area is usually from 8" x 10" to 20" x 25", and digital cameras can also be used for oversize material, but they are an expensive option compared to flatbed scanners.

Bound pages that cannot be disbound, and pages in bindings that cannot open fully to 180 degrees require flatbed scanners with a right angle, prism, or overhead capture array. Digital cameras, with appropriate easels, book rests and weights are a versatile option for bound material. Camera beds or mounts, lighting, lenses, and filters all add to the cost and complication but make digital cameras more versatile tools for capturing manuscripts, bound volumes, original works of art, prints, out-size material and artifacts.

To achieve the highest quality scans of transparent media (e.g. 35mm slides and negatives, 6x4 and large format transparencies and microfilm) specialist equipment such as slide and film scanners, microfilm scanners or drum scanners should be used. Some flatbed scanners, with a dual light source, can handle transparent media though they often lack the dynamic range comparable to that supported by transparency scanners. However, you will not achieve as high a quality image as you would with a dedicated film or slide scanner. These have an inherently higher resolution, appropriate for the small size of the original, hold the transparencies more closely and securely, and frequently have negative color compensation to correct color casts for different types of film.

Audio and moving image materials present their own problems for digital capture. Not only is there a variety of source formats, including wax cylinders, 33, 45 and 78 rpm records, 8-track and cassette tapes, two-inch and VHS video in PAL and NTSC formats, but it is often very difficult to obtain access to analog devices for playback and linkage is difficult.

Definition Box:

Audio-Visual Facilities:

- Audio capture card required for sound material or video capture card required for moving images
- Source devices, such as 78rpm record players and tape players.
- Mechanism for connecting these devices digitization equipment
- Intermediary device, such as a DAT (capable of handling ASEBU and SPDIF digital audio) machine

Link Box:

There are a number of audio and video digitization projects that are just getting started:

RAI: http://www.rai.it/portale

BRAVA: Broadcast Restoration of Archives through Video Analysis
http://www.ina.fr/recherche/projets/encours/brava/

COLLATE: Collaboratory for Annotation, Indexing and Retrieval of Digitized Historical Archive Material http://www.collate.de/index.htm

PRESTO: Preservation Technology for European Broadcast Archives
http://presto.joanneum.ac. at/index.asp

AMICITIA: Asset Management Integration of Cultural Heritage In The Interchange between Archives
http://www.amicitia-project.de/ami_home.html

The 3D representation of objects, from coins to buildings, is at the forefront of current digitization developments. At present the technology can be divided into two broad categories. The first, and simplest, is to create a moving image of an object. This is achieved by moving a digital camera around the object, or rotating the object in front of a fixed camera, while taking a series of still images. These images are then compiled to create a moving image of the object. The most common format for this is QuickTime VR. This is a reliable technology that requires a digital camera and mount or turntable. However, it does not provide a true 3D representation of the object because while only two planes are captured and displayed, it still represents 3D objects using two spatial planes. The viewer cannot manipulate the object, and the views provided are fixed and pre-determined.

Creating a true 3D representation of an object requires that the dimensions and features of the object be modeled. That is, the three dimensions of the object are represented in the computer as a set of coordinates. Attached to this "frame" are the textures of the object to provide the surface details. At present most 3D imaging technology remains in the sphere of industry. The technologies used to capture coordinates, render the model, and interact with the 3D representation (such as haptic feedback systems that allow one to "touch" the object, or 3D printing to create facsimiles) are often quite costly and require a relatively enormous amount of computing processor power compared to the average desktop computer (in 2002). As such, 3D modeling devices remain application-specific, for example body imaging, prototyping or CAD/CAM applications. However, it was not long ago that digital imaging was the sole preserve of medical applications. During the next ten years we should see increasingly cost-effective and user-friendly devices that will bring 3D modeling into the mainstream.

Definition Box:

Virtual Reality:

Virtual reality can be described as an interactive, self directed, multi-sensory, computer generated experience which gives the user an illusion of participating in a three dimensional environment, even if a synthetic one. For cultural and heritage institutions, this may mean using virtual reality to create virtual representations of three dimensional objects in their collections or to create representations of environments, such as an Egyptian tomb, an ancient Persian palace, a historic Greek theatre or an ancient landscape. These three-dimensional objects could range from coins, vases, and sculptures to representations of whole rooms of collections.

Metadata

Metadata is an indispensable part of any responsible digitization program, and considerable attention has been paid to the definition of high-quality metadata standards for various purposes. (The appendix on metadata provides more detail on different types of metadata, and on specific metadata schemes and their uses.)The availability of accurate metadata is as important as the digital surrogates themselves for accessibility,

usability and effective asset management. In many instances institutions will already have substantial metadata about the analog object (for instance, catalog records) much of which can be applied to the digital object. The project will be able to reduce its metadata creation costs by building on existing metadata. When selecting material for digitization you may wish to give priority to material for which partial metadata already exists.

It is crucial to remember to determine the status of the existing metadata, when you are assessing resource requirements. In an ideal world the existing catalog or finding aid would be complete and up to date. However, many libraries, archives and museums have a backlog of cataloging work, and part of a collection selected for digitization could fall into this category. Therefore, it may be necessary to devote time to locating missing information for your metadata records. You must then decide whether to seek information just for those fields required for the metadata, or to update the original catalog record in its entirety. Digitization provides an economical opportunity for institutions to expand their metadata, so consider the possibility of seeking extra funds or devoting more resources to this activity. Some of the new elements required for the metadata record of the digital object can be generated automatically: for instance, automatic metadata creation is a feature of much high-end digital camera software and of some OCR systems. Alternatively, a project may need to develop its own system, and can greatly improve the efficiency and accuracy of technical metadata. There is a general dearth of metadata tools, which poses a problem for the efficient creation and management of metadata for many projects. There is therefore likely to be a significant element of manual work, whether this lies in adding digital objects to existing electronic catalogs, creating records for web-based delivery such as Dublin Core, or implementing encoded metadata schemes such as EAD. Creating a metadata record will usually take as long as creating the digital surrogate and if detailed encoding schemes such as Encoded Archival Description or Text Encoding Initiative are used, this process can be considerably longer.

METADATA RESOURCES:

GENERAL METADATA RESOURCES

1. Canadian Heritage Information Network Standards Page: http: //www.chin.gc.ca/English/Standards/metadata_intro.html
2. J. Paul Getty Trust, Introduction to Metadata: http://www.getty.edu/research/institute/standards/intrometadata/
3. Extensible Markup Language: http://www.w3.org/XML/
4. International Federation of Library Associations and institutions. Digital Libraries: Metadata Resources: http://www.ifla.org/II/metadata.htm
5. Text Encoding Initiative: http://www.tei-c.org
6. Metadata Encoding and Transmission Standard (METS): http://www.loc.gov/standards/mets/

***METADATA MENTIONED ELSEWHERE IN THE* GUIDE**

- Section III: Selecting Materials: Metadata & Interoperability. The Dublin Core metadata initiative http://dublincore.org/

- Section IV: Rights Management: Technologies for Copyright Management and Protection.
  - The Open Digital Rights Language Initiative (ODRL): http://odrl.net/
  - Digital Object Identifier (DOI): http://www.doi.org

- Section V: Digitization and Encoding of Text - Text markup schema. Text Encoding Initiative (TEI): http://www.tei-c.org

- Section VI: Images
  - Descriptive:
    - Library of Congress Subject Headings (LCSH): http://lcweb.loc.gov/cds/lcsh.html#lcsh20
    - Categories for the Description of Works of Art (CDWA): http://www.getty.edu/research/institute/standards/cdwa/
    - Art and Architecture Thesaurus (AAT) http://www.getty.edu/research/tools/vocabulary/aat/about.html
    - VRA Core Categories: http://www.vraweb.org/vracore3.htm
    - Dublin Core Metadata Element Set: http://dublincore.org/documents/dces/

  - Structural:
    - Synchronized Multimedia Integration Language (SMIL) http://www.w3.org/AudioVideo/
    - Metadata Encoding and Transmission (METS) Standard: http://www.loc.gov/mets

  - Administrative:
    - A Web Hub for Developing Administrative Metadata for Electronic Resource Management http://www.library.cornell.edu/cts/elicensestudy/
    - Digital Library Federation, "Structural, technical, and administrative metadata standards. A discussion document:" http://www.diglib.org/standards/stamdframe.htm

- Section VII: Audio and Video Capture and Management
  - Dublin Core Metadata Implementers: http://www.fiu.edu/~diglib/DC/impPurpose.html
  - Synchronized Multimedia Integration Language (SMIL) http://www.w3.org/AudioVideo/
  - Metadata Encoding and Transmission (METS) Standard: http://www.loc.gov/mets
  - MPEG-7: http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm
  - Authority Tools for Audio-Visual Catalogers: http://ublib.buffalo.edu/libraries/units/cts/olac/capc/authtools.html#g
  - Authority Resources for Cataloging Popular Music: http://www.music.indiana.edu/tech_s/mla/wgpms/wgpms.htm
  - Library of Congress's Digital Audio-Visual Preservation Prototyping Project: http://lcweb.loc.gov/rr/mopic/avprot/avlcdocs.html#md
  - Library of Congress's Digital Audio-Visual Extensions to METS Standard http://www.loc.gov/rr/mopic/avprot/metsmenu2.html
  - Cinemedia's SWIFT project for on-demand delivery of film and video: http://www.cinemedia.net/SWIFT/project.html

- Section VIII: Quality Control and Assurance: Importance of Quality Control and Assurance of Metadata

- Section X: Distribution:
  Metadata Harvesting
  - Clifford Lynch, "Metadata Harvesting and the Open Archives Initiative," *ARL Bimonthly Report* 217 (August 2001): http://www.arl.org/newsltr/217/mhp.html
  - Donald Waters, "The Metadata Harvesting Initiative of the Mellon Foundation," *ARL Bimonthly Report* 217 (August 2001): http://www.arl.org/newsltr/217/waters.html
  - The OAI MHP protocol: http://www.openarchives.org/OAI_protocol/openarchivesprotocol.html
  - MHP tutorial: http://library.cern.ch/HEPLW/4/papers/3/
  - CIMI Working Group: http://www.cimi.org/wg/metadata/
  - CLIR Metadata harvesting project: http://www.clir.org/activities/details/metadata-docs.html
  - DLF and Metadata Harvesting: http://www.diglib.org/architectures/mdharvest.htm
  - University of Illinois at Urbana-Champaign Metadata Harvesting services: http://oai.grainger.uiuc.edu/

- Section XIII: Digital Asset Management: "Metadata definition and management"

- Section XIV: Preservation:
  Institutional Approaches
  - METS (Metadata Encoding and Transmission Standard) http://www.loc.gov/standards/mets/
  - UK Cedars project: structure for preservation metadata: http://www.leeds.ac.uk/cedars/metadata.html
  - Australian Preserving Access to Digital Information (PADI): Research Overview and Updates on Preservation Metadata: http://www.nla.gov.au/padi/topics/32.html
  - NISO: Technical Metadata for Digital Still Images: http://www.niso.org/standards/resources/Z39_87_trial_use.pdf
  - OCLC/RLG Preservation Metadata Working Group: http://www.oclc.org/research/pmwg/
  - Reference Model for an Open Archival Information System (OAIS): http://www.ccsds.documents/pdf/CCSDS-650.0-R-2.pdf

Project Management

Many different approaches to managing projects are possible. While we found little evidence of the conscious adoption of a project management model, such as PRINCE 2 (http://www.kay-uk.com/ prince/princepm.htm), most projects implemented many of the key features of successful project management. As understanding of digitization becomes more commonplace it may not be necessary to "hot house" prototype projects in the manner that many early projects experienced. However, it should also be recognized that integrating existing projects into host institutions often adds a layer of bureaucracy.

The Genealogical Society of Utah provides a good example of a comprehensive project management model. Each imaging project undertaken follows six stages:

1. Negotiation and project administration
2. Capture Convert Acquire
3. Image and metadata processing

4. Storage and preservation
5. Indexing and cataloging
6. Access and distribution

All projects will need to consider these six areas in setting up their own project management systems.

You do not necessarily need to adopt all the activities of a project management methodology; rather you need to scale the method to the needs of your project. The whole process should be determined by the project's objectives and rationale for creating the digital deliverable. Each process should be defined, together with the specific objectives to be achieved and activities to be carried out. The various roles and responsibilities should be detailed (defining job descriptions and breaking finances down aid in this — see above) and adapted to the size and complexity of the project. This should enable the efficient control of resources and facilitate regular progress monitoring. Regular reviews should be used to ensure that the project's objectives, which may change during the project lifecycle, are being met. Whatever project management method is adopted, it should provide a common framework and delineate milestones for all elements of the project.

In summary, your project management methodology should make possible:

- The use of pilot projects and feasibility studies to shape the overall scheme of activity
- Controlled and organized stages
- The establishment of a project plan with milestones
- Regular reviews of progress against plan and against objectives
- Control of any deviations from the plan
- The involvement of all constituents at the right time and place during the project
- Good communication channels between all constituents in the project and the sponsoring institution/s

Other key features are the need for one project manager to have ultimate responsibility and for the project advisory group to provide management quality control and assurance. In distributed projects, site managers are recommended in addition to an overall project manager. Most projects have relied on internal project management expertise, supplemented by external advice. Although many projects started as relatively autonomous there is a clear trend for project management structures and the project organization to be integrated into the host institution's structure. This may be a natural progression for projects as they mature, but new projects may consider whether they should adopt it immediately.

Work flow and costings

While few of the projects interviewed carried out benchmarking tests most had conducted pilot studies. These were undertaken for a variety of reasons:

- Technical feasibility
- Technical forecasting
- Workflow analysis
- Training needs

When considering technical forecasting or prototyping, particularly in relation to costs, remember that there may be no corresponding benefit, and if there is a benefit it will vary for different types of content. Few projects in the humanities and cultural sector charge users for the digital deliverables. As such the cost/benefit may simply be realized by the ability of the project to amortize the depreciation on the equipment. A new high-resolution camera may pay dividends for fine textual or line art material, but not so for color images. Similarly, a device that enables the digitization of material that previously could not be captured, such as a 3D modeler, may not make financial sense if a project has to build in a profit or depreciation margin. However, if the device makes an important collection more widely available, the public access benefit may outweigh the financial costs.

Where any form of pilot study is undertaken it is important to build this into the project design and development cycle. For example, the University of Virginia Library's Special Collections department delineates its project work as intricately as possible before extrapolating its workflow and costings. This has given the project reliable data to forecast costs, but there are some areas where measurement has proved inaccurate, such as network transfer rates. The UVA Special Collections department also has a scheduling calendar tied to a tracking database to generate quality control and assurance checks and back-ups. In this respect it is typical of the projects surveyed which all use flowcharts, spreadsheets or Gantt charts to plan and monitor their workflow and costs.

If you are considering using a cost model (see above), it is important to include all the relevant costs, not just the obvious items such as equipment and staff time. You will also need to decide on what basis to evaluate — for example, costs per unit to be digitized or costs per hour. The table below provides a checklist of the factors that should be built into a cost model.

Finally, one further area to be aware of as you develop your cost estimates is digital asset management. In digitizing an image collection, for instance, you may well be generating a number of different kinds of digital objects-archival masters, delivery masters, thumbnails and other deliverables-which in turn will require storage, tracking, documentation, and upkeep. This process may require a significant commitment of

resources and will need to be planned carefully. Section XIII covers digital asset management in detail.

Cost Model Factors	
Equipment	Purchase Maintenance Repair
Software	Purchase Upgrades
Staff	Salary (including benefits and insurance,) Training Recruitment Travel & subsistence
Utilities	Heat Light Water Phone Postage
Building	Rates Maintenance Upgrading/expansion

Analysis Box:

Costs of Digitization Programs

There is little information available about costs and this is an area where future work is necessary. In a rare exception, Steven Puglia analyzed the costs of digitization programs, in particular from the Library of Congress Ameritech competition and the National Archives and Records Administration's Electronic Access Report. The costs discussed are mostly *projected* and *estimated* costs, a problem discussed in the conclusion, which suggests that further studies are necessary. After an initial discussion on general costs of projects—it appears that on average, a third of the costs incurred by projects is the digital conversion, slightly less than a third is metadata creation, and slightly more than a third is made up of administrative and quality assurance tasks—the emphasis turns towards long term maintenance costs. The author suggests that these are not often taken account of with the project costs.

Three types of maintenance of digital objects are considered, each with mounting costs in relation to the initial costs per image:

- The first projection is an example taken from the maintenance of NARA objects. This involves only keeping one copy of master files and offering thumbnails and access online. The cost of this basic maintenance is in the region of 14-38% of the initial imaging cost.
- The second projection comes from Cornell report on computer output microfilm. Applying this to NARA, the cost would be 55-60% of the initial cost per image.
- The third projection looks at outsourcing to a commercial firm. This would cost 275-796% of initial cost. But it must be remembered that the other two costs do not include major IT infrastructure costs and thus are false figures, whereas the private firm costing will include this.

In conclusion, it is suggested that digital imaging may not be the best approach for long-term retention of information. Institutions can only justify retention if the images are used. Analog retention is the best way of holding materials in the long-term. In addition, it would be instructive to use figures from final project costs and also examine costs per person and production per person.

For the full report see: Steven Puglia, 'The Costs of Digital Imaging Projects', RLG DigiNews, October 15 1999, Vol. 3 No. 5. http://www.rlg.org/preserv/diginews/diginews3-5.html

Conclusion

At the start of any project, project planning feels like a way to exert control, eliminate risk, and guarantee a successful outcome. Certainly without good planning, the likelihood of failure and inefficiency is much greater. But you can be a better project planner by recognizing that the goal is not to eliminate risk but to prepare for it—not to control every variable but to create a project framework within which your team's response to the unforeseen will be resourceful and effective. In the technology domain, change and unpredictability are facts of life, and often represent opportunities rather than disasters for a well-planned project. Your planning goal should be to create a flexible, adaptable system whose staff and procedures can accommodate change. Your aim as a project leader should be to distinguish between what is essential—the central project objectives, the strategic components that will ensure long-term viability—and what is merely instrumental detail.

III. Selecting Materials: An Iterative Process

Introduction

In this section we look at the issues you need to examine when selecting material - whether selecting physical originals for digitization, or reviewing born digital materials for preservation or republication. We also show how you can ensure that this process takes into account the aims and characteristics of your organization, the profile and needs of your users, and the characteristics of your collections. Some central questions you'll need to consider include the following:

What are the aims of the digitization project or program and how can these guide selection, in cases where limited resources prevent you from digitizing the whole collection?

- How do these issues apply to material created in digital form?
    - Who will be using these resources?
    - What are their needs?
    - Do the collections you selected for digitization need special handling?
    - What are their characteristics and how do these affect the digitization process?

These are only some of the questions you will need to address when embarking upon a digitization project.

Characteristics of original material

Strategic knowledge of institutional holdings

Collections are a vital component of the intellectual capital of cultural heritage institutions and in most cases, their *raison d'être*. Successful digitization programs start from a strategic knowledge of the institution's collections and their relation to the institutional mission. A strategic assessment of this sort is also a vital step for regional, national, and international initiatives and collaborative or larger programs. Unfortunately, this prior analysis of existing holdings was often omitted in many of the early digitization projects.

Planning for digitization should start from a study of the analog sources, the physical materials themselves, rather than in response to available technology or other pressures.

The analysis should include an assessment of the physical objects, their condition, characteristics, and educational, cultural, historical, and aesthetic value (Ross 1999).

Planning for digitization and digital publication should start from a study of the analog sources, the physical materials themselves, or, in the case of born digital material, the digital collections, rather than in response to the nature of the available technology or other pressures. As we are now passing from the first stage of experimental digitization projects to sustainable long-term programs, the technology itself holds fewer mysteries and attractions, and problems of scale and long-term planning come to the fore. It is thus important to carry out an overall assessment of institutional holdings before deciding on digitization priorities. In this way, even if digitization starts at a smaller scale with potential for expansion, its development will be well planned and will complement the institution's strategy and objectives.

Where strategic knowledge of the institutional holdings does not already exist, this type of assessment will require resources if it is to be carried out in a systematic and thorough way. It needs to be planned ahead, have institutional support at all levels, and include all team players. This can be demanding in terms of time and staff; however, the required resources will be well spent, as this analysis will be a valuable tool in planning all the organization's activities. It will clarify the strengths of the collection and will place them in a local, regional, national and international context.

When planning an evaluation of this type, physical access to the original materials is a fundamental consideration. This type of assessment needs to include both primary and secondary material and to examine the condition and size of both. A systematic assessment of the institution's collections will complement the resource inventory, as well as the institutional digitization policy we mentioned in Section II, highlighting particular strengths and resources in answer to the questions 'Who are you?' and 'What do you have?'

Intellectual value

For cultural institutions that are in the process of examining their assets, it is important to first assess their intellectual value to the intended audience. The project team, which should include custodians of the material, should assess whether the analog materials have sufficient intrinsic value and quality to ensure interest in their digital surrogates and to sustain the levels of access made possible by digitization. The Question Box below contains examples of the kinds of questions you may wish to ask when considering whether material should be selected for digitization.

Question Box:

Questions to Ask Before Selecting Material for Digitization:

- Is the information content of the analog objects high?
- What is the intellectual significance of the analog materials?
- How do they compare in terms of importance, authority, uniqueness, and timeliness?
- Is the object you are considering for digitization a good example of the period under examination?
- Is it original?
- Are there alternatives in the collection more capable of illustrating particular points?
- Is it complete or are there any parts missing?
- Is it up to date?
- Is it accurate?
- Would it withstand the test of time or is it only of ephemeral interest today? (Is it, for example, an important testimony of the past, of another culture or our own, of an artistic movement, an author, or a scientific discovery?)

As is clear from these questions, there is an element of subjectivity when making these judgments and weighing the intellectual nature of the various collections. Our perceptions and evaluation depend on our perspective and are subject to change. It is therefore advisable to consult widely within your institution and peer group as well as with your users in order to reach a general consensus in your decisions. Establishing the needs of potential users (see Section XII, User Evaluation), and determining what collections in other institutions might be digitized (see Section IX, Working with Others), will further enhance this activity, and can also help inform the institutional collection policy overall.

Added value and functionality

While examining the current intellectual value of the original materials, it is also worth considering the advantages of digitization in this area. A good example is the digitization of the Anglo-Saxon Beowulf manuscript, which was carried out by the British Library in collaboration with Professor Kevin Kiernan from the University of Kentucky with the close collaboration of scholars, curators, conservators, photographers, and technical experts (Kiernan 1994; Prescott 1998). This is one of the greatest treasures of the Library and of paramount importance for scholars of Old English. The manuscript had been badly damaged in a fire in 1731 and in order to protect the brittle and smoke-stained pages, each one was mounted in a protective paper frame in the mid-nineteenth century. In order to have a retaining edge for the frame, the letters around the edge of the verso of each leaf were covered, obscuring hundreds of letters from view. The digitization of the manuscript

was one of the Library's early worldwide digitization projects, conceived in 1992 and completed with the production of a CD-ROM in 1997. Using a high-end Kontron digital camera (manufactured originally for medical imaging), while lighting each page with fiber-optic lighting, it was possible to capture digital images of the pages where the hidden letters were revealed.

In this case, digitization offered tremendous added value to the original manuscript, bearing in mind that some of those hidden letters represent the only known record of some Old English words. The project subsequently expanded to include a collection of digital images of all the primary evidence of the Beowulf text held in different geographic locations, ranging from the Royal Library in Copenhagen to Houghton Library at Harvard University, thereby creating a new resource and tool that allows enhanced study of the manuscript.

Digitization can also enhance the intellectual value of your original collections by allowing new possibilities for education and access, for example:

- by providing good quality digital surrogates of art works that are dispersed in different institutions around the globe;
- by allowing the creation of personalized study paths through the material, or juxtapositions of the works according to themes or particular interests;
- by bringing the collections to new audiences and allowing new ways of exploring and enjoying them;
- by offering the ability to search and manipulate text in electronic form, allowing a variety of ways to interrogate the material;
- by offering access to different versions and editions, opening up a richer and more complex experience of the text than is possible in printed form;
- by creating powerful electronic indexes which allow quick searching of video or audio clips and their synthesis in new forms.

Physical properties of source material

While the intellectual significance and content of your materials are very important, their physical characteristics also influence selection for digitization since they directly affect the digital outcome. Therefore, the analysis of the physical characteristics of the collections is an important step that will define how to handle the material, while deciding the subsequent digitization process. The custodians of the materials should ideally consult with those responsible for the digitization program to decide on what information is relevant. We include here some suggestions that are intended as a starting point to guide you in this process of examining the material itself and recording information about it. You might want to devise your own categories or expand on those listed in the Checklist Box: Physical Properties of Source Material.

Question Box:

Physical Properties of Source Material

- Type and category of object
  (e.g. Is it a book, manuscript, photograph, sound recording, TV broadcast?)
- Production process
  (e.g. Printed document, handwritten text, engraving, woodcut, wax cylinder recording, recording in mono or stereo? Is it an original or a reproduction/intermediary?)
- Date
  (e.g. How old is it? Do you have information about when it was made? If not, can you find out or estimate?)
- Physical size and dimensions
  (e.g. Is it a book with pages of regular (letter) size? Is it uniform? What is its length in cm, inches, duration of an audio-visual tape in hours/minutes/seconds, number of reels of film width; depth; thickness; weight?)
- Media Type
  (e.g. Paper, leather, wood, magnetic videotape, vinyl record; combination of materials; gold leaf details in manuscripts?)
- Format
  (e.g. 78 rpm disc, wax cylinder, reel-to-reel analog tape recording, DAT (Digital Audio Tape), Betacam SP tape, NTSC or PAL format video recording?)
- Sensitivity to light
  (e.g. What kind of lighting levels can it be safely exposed to during digitization? For how long?)
- Color information
  (e.g. Does it contain color information? Does color convey important information in this case? Is it an important element for the understanding and appreciation of the object?)
- Tonal Range
  (e.g. Does it have a wide tonal range? Is this an important element for the understanding and appreciation of the object/recording?)
- Noise
  (e.g. Does the audio recording have audio hiss, clicks and pops? Are there background sounds or images, which were captured in the original sound or moving image recording, that are not related to the main material? Is it important to preserve these?)
- Characteristics of born digital material
  (e.g. File format, resolution or sampling rate, bit-depth or rate, compression method, file size?)
- Characteristics and structure of informational content
  (e.g. For printed documents: does it include both illustrations and plain text? For sound recordings: what is the duration of the songs recorded and how many are included in the tape? For video recordings: what is the duration of the film recorded?)
- Structure of the material
  (e.g. Is the material bound or mounted?)
- Condition of the material and conservation
  (e.g. What is its state of preservation? Has it been assessed by conservators? Should it be conserved? Does it require any special handling?)

Digitization aims

In addition to the steps already outlined (examining the intellectual value of the materials, how this could be enhanced by digitization, and taking into account the physical characteristics of the collection), guiding principles for the selection process are the aims of the digitization program itself. These will vary between institutions, but we include

here some of the most common general aims for starting a digitization program and how they might affect the selection of the material. In order to prioritize the selection process, examine which collections would be good candidates for:

- improving access;
- assisting the preservation of the original by reducing its wear and tear;
- supporting research activities;
- meeting user needs;
- allowing the use of good quality existing metadata;
- complementing other digitization efforts, either within the institution or beyond, without creating any rights management problems.

Increased access

In the examples of the Beowulf manuscript and the digitized art works previously cited, it is obvious how digitization can significantly increase access to resources. Examining whether digitization would indeed substantially increase resource accessibility is another criterion that can guide you in the selection process. This might involve asking the following questions:

- Does your institution already provide access to parts of the collection in non-digital forms?
- Is this easy and adequate?
- Does it serve your user community sufficiently? (This is further explored later in this Section — see User Needs and Demands.)
- Are there limitations to access due to the condition of some of the original collections?
- Are some of the original materials dispersed geographically within your institution or beyond?

The answers to these questions will reveal good candidates for digitization or will at least assist in establishing priorities. For example, some of the easily accessible analog material might move further down the selection list in favor of materials that are in remote stacks or storage areas. On the other hand, you might still select easily accessible materials, if they are so important to your users that the current level of analog access that you provide is inadequate. Other candidates might be significant but under-utilized collections with high intellectual value and relevance to the interests of your users. In this case it is worth further examining the reasons behind their limited use. Could you do anything to remedy the situation using methods more affordable than digitization?

When trying to identify a user group that would benefit from access to digitized materials, most institutions start from information collected about the current use of analog materials. The quantity, information needs, characteristics and location of current users can be a useful starting point for estimating future use. However, this is not always accurate, as digitization can increase access to and use of material that was hitherto unknown or underused. Additionally, the potential of digital information and web access can be so powerful and difficult to predict, that the actual users of digital resources are not always the same as those anticipated, especially where institutions are sharing digital surrogates of their assets with a worldwide community.

Preservation

Digitization of analog materials is not a substitute for investment in their conservation and preservation, but can assist the preservation of the original. Heavily used materials can benefit from the creation of faithful digital copies to prevent the deterioration of the originals, as can materials that are fragile or at risk. In this case, you should assess whether the benefit of digitization is greater than the risk placed on the material during the process of digitization. Digitization should also be a priority for cases where the existing storage medium is no longer suitable, as, for example, with nitrate film. Digital files are themselves vulnerable to obsolescence, chiefly as a result of changing file formats, but also through deterioration of the storage medium. There are a number of strategies for overcoming this problem: careful specification of settings for quality capture and regular quality assessment; consistent use of well-defined and detailed metadata; and use of widely recognized, standard formats. These are discussed in more detail in Section XIV on preservation.

Research

Support for research activities is often an important incentive for digitization, although academics need to be aware that the resulting digital resources may not remain accessible indefinitely (see Section XIV: Preservation). Digitizing high quality, unique and original material can improve access for a wide community of researchers at all levels. It can also enable interdisciplinary collaboration and study from a wider range of scholars than was previously possible. For example, the creation of the Perseus database with its associated tools and resources has encouraged the study of ancient Greek and Latin culture in new ways, combining the analysis of texts, material culture, art, architecture, history, and geography.

> *Interface design and appropriate metadata ... [require] considerable intellectual effort, knowledge, and resources to bring intelligence and context to the individual digital files, regardless of their quality.*

While access to digital surrogates will never supersede the need for researchers to use the original objects, digitization offers added functionality that can assist research in other ways. It has some important shortcomings: it takes away the immediacy of the original and its impression of size and color (which are still not displayed accurately with the existing technology today), and it eliminates or obscures some of the original context. The program team that selects and prepares materials for digitization can counterbalance these effects by providing sufficient context and accompanying information to make the digital objects meaningful. For example, good interface design and appropriate metadata can ensure that digitized illustrations do not appear on the screen in isolation, without reference to the rest of the text or the other related volumes; or that early LP recordings will not be played without reference to the text and images on the original disk jacket. Considerable intellectual effort, knowledge, and resources are required to bring intelligence and context to the individual digital files, regardless of their quality, which are little more than an electronic likeness of the original object. The program team must address these issues: it must examine the type and level of context it wants to provide, which should reflect the available resources as well as the research needs the project aims to support.

We have only begun to explore the possibilities of research tools for manipulating, searching, and studying digital resources. The ability to carry out sophisticated searches and accurate retrieval through rich materials is assisting all areas of research. In the area of text retrieval and processing, for example, the availability of large electronic corpora has contributed significantly to fields such as lexicography, socio-linguistics and authorship attribution studies. Developments in automatic translation programs and multilingual thesauri will allow much wider use of hitherto unknown resources and will enable cross-lingual searching. Image retrieval presents greater challenges. At the moment, few effective content-based image retrieval (CBIR) products have yet reached the marketplace. Most CBIR tools retrieve images based on appearance and the numerical characteristics of color, texture, and shape, rather than intellectual content and image semantics (Lesk 1998), and even in these areas they produce poor quality search results. However, CBIR continues to be the focus of intensive academic and commercial research activity. Even with the weak experimental applications that are available, including IBM's QBIC, Virage's VIR Image Engine, and Excalibur's Visual RetrievalWare, these tools clearly hold considerable promise to improve researchers' access to digital resources. Although there is still no substitute for trained catalogers, librarians, and subject specialists tagging the images with keywords, this is an area where future developments might revolutionize the use of images.

Link Box:

CBIR Tools to Watch:

IBM's QBIC: http://wwwqbic.almaden.ibm.com/

VIR Image Engine: http://www.aa-lab.cs.uu.nl/cbirsurvey/cbir-survey/node41.html

Excalibur's Visual RetrievalWare: http://www.excalib.com/

A wider and richer collection of materials is also becoming available for research through advancements in even more challenging areas: development of next generation scanners for digitizing three-dimensional objects; the use of speech recognition for the indexing and retrieval of digitized speech archives; and research into video analysis and indexing systems that can parse and analyze hours of video, identify events as they occur, extract embedded textual data, employ continuous-speech recognition to convert spoken words into text and convert all this information into a searchable database. Collaboration between institutions (discussed in Section IX) also has a vital role to play in answer to the common current complaint that resources are not sufficiently large and comprehensive.

The digital environment continues to suffer from a number of limitations and restrictions that have a subsequent impact on research. For example, a large percentage of the material digitized thus far by institutions in the cultural and educational sector, belongs in the public domain. The selection has been driven more by copyright restrictions and less by the material's intellectual significance and usefulness for research and education. The rights to moving image and recorded sound material can restrict access to some information sources. Although OCR accuracy has been continuously improving for the last few years, the poor results with non-Latin characters has meant that very few texts in non-Western languages are available online (Smith 1999). Keyboarding is an alternative in these situations — not only for non-Latin alphabets but also for manuscript materials where OCR is not an option in any case. Several projects (for instance, the Japanese Text Initiative at the University of Virginia Etext Center, http://etext.lib.virginia.edu/japanese/) are creating resources in this way and using Unicode to handle the character set encoding. However, the added cost of these methods necessarily limits their size and scope. These issues go beyond technological development and are associated with socioeconomic conditions and pressures; they will need to be addressed by the research community as they have serious implications for the availability of resources and the direction of scholarship.

User needs and demands

User requirements should also guide the digitization selection process. Although it is difficult to predict future usage accurately in the electronic environment, current usage patterns of your collections can provide valuable pointers. Knowing the profile of your users and the ways they access information can highlight heavily used areas, limitations, and possibilities for development. Consider the following: What is the size of your user group? How are they distributed geographically? Is there a need for decentralized access of resources, e.g. from different institutions or from home? Will your users need special tools and facilities for using the digital resources? Do you have several different groups of users who may need different facilities and levels of access? If you do not already have sufficient information about your users to answer these kinds of questions, you will need a strategy for collecting that information and assessing whether digital resources can answer existing and future user demands. Digitization can also be used to bring in new types of audiences and help to open up specialized resources to a wider public. Digital collections can support lifelong education and learning, and they can also promote social and cultural inclusivity by opening up access for socially disadvantaged groups. If

providing access to new user groups is one of the aims of the digitization program, it is important to try to involve the targeted users early on in this process. Consultation with these groups can bring new perspectives to your collections, assist in the selection of materials and provide information about new ways of using it. For example, providing schoolchildren with access to digitized resources from museums and galleries can be facilitated by early involvement with targeted user groups and discussions with schoolteachers about the ways that these can be used in the classroom. Working with different communities and ethnic groups can result in a variety of different ways of looking at the same material and making associations and links that go beyond the Western European, white middle-class biases.

In order to provide digital resources that are relevant and useful for your target groups, you will need to address the need for further evaluation and research. Evaluation involves user needs analysis and assessment using some of the methods we discuss in Section XII on user evaluation. This can take place before a digitization program begins, or before developing an existing program further. Information about how digital collections are being used is very limited at present and includes very little qualitative, in-depth detail that goes beyond the standard web usage statistics. Although we know, for example, that the American Memory web pages are used by millions of users every year[1], we know very little about how these resources are being used by a variety of audiences, even though this was one of the few projects that included an early evaluation survey of its pilot in 1991-3 with various schools, colleges, universities, and special, public, and state libraries (American Memory Evaluation Team 1993).

The level of technology available to user groups for accessing the digital material is another important consideration when selecting materials and approaches for digitization. If users do not have sufficient bandwidth to download or view large still or moving image files, for example, this material cannot be successfully disseminated. On the other hand, the digital environment encourages new expectations from users that your organization might not always be ready to meet. You will need not only to decide which are the most important needs to meet, but also to develop a clear articulation of the priorities and strategies that led to these choices.

Intellectual property and rights (IPR) management

As soon as you have identified appropriate material for digitization, you should tackle the intellectual property issues, since securing permission for materials you do not own can be a lengthy and often a costly process. If the materials you want to digitize are in the public domain, or if you own the copyright and control the digitization rights, then you can probably proceed without hindrance. Bear in mind that these issues—even concerning materials in your own collection—can be complex, for instance in cases where a photographer has been employed to take photographs of objects in the collection and these photographs are themselves to be digitized, or where gifts or bequests are concerned.

If the materials you want to digitize are not in the public domain, and if you do not control the copyright, you will need to identify the copyright holder and seek permission to digitize and publish the materials in question. You may already know who the copyright holder is, and may already have permission to use the materials for certain purposes, but be sure that this permission covers digitization and digital publication. Bear in mind as well that you may be asked to pay a fee or royalties for digitization or publication rights. In some cases, your needs may be covered by fair use (see more about fair use in Section IV on Rights Management) in which case specific permission from the copyright holder is not required.

It is possible that you may not be able to identify the copyright holders, or that they may not respond to your inquiry. If permission cannot be obtained even through good faith efforts, you need to proceed—if at all—very cautiously and assess the risks carefully. Again, in such cases, fair use may justify your inclusion of the materials in question. Restricting access to the materials may also limit your risk. Section IV includes more information on risk management, as well as links to resources on IPR.

Be very cautious, however, when considering publishing material on an unrestricted website. If what you plan to digitize cannot be justified as fair use, and if you cannot secure permission for digitization and electronic distribution of digital copies, you need to assess very carefully the risks of proceeding.

If, after assessing copyright status and fair use considerations, your institution finds that permission is required, you need to make a careful assessment of the value and costs that are at stake. Securing permission can be an arduous, time-consuming, and costly process, often with an uncertain outcome in cases where the copyright holder is either unknown or unwilling to give permission. The costs and uncertainty are frequently reasons why institutions decide not to digitize a particular work, despite its desirability. As indicated above, this raises the concern that some cultural materials will remain unavailable simply because their copyright status is uncertain or heavily restricted. Similarly, these materials may be available in digital form but only through commercial products, if copyright holders require fees that are beyond the reach of non-profit digitization efforts. Some have argued that cultural institutions with rich collections should avoid the effort and cost of securing permission to digitize materials they do not own, focusing instead on materials which they do control. (Zorich 1999), While this approach makes practical sense, it means that the shape and scope of digital collections will tend to reproduce the boundaries of the individual collection, rather than bringing together all related materials in a given domain, regardless of location and ownership. Collaboration can play a very important role in overcoming this limitation, but requires careful negotiation and coordination between institutions (see Section IX, Working with Others).
Even in the cases where your institution has been granted copyright to the analog material, there are further issues to be considered. You should examine, first, whether the rights you have actually extend to the digitization and distribution of electronic copies, particularly in the cases of donations and bequests. Even when there are good reasons for using materials to which you do not own copyright, you should be careful about the permissions or licensing agreements you negotiate and ensure that you have taken proper

legal advice. Museums and art institutions should also be careful to protect the rights of the artists whom they display. Furthermore, when commissioning photographic documentation of your collections, you should ensure that the photographer's contract takes into account the digital environment and that you have cleared the rights for digitization and distribution of digital copies.

When you own the copyright of the materials you are selecting for digitization, make sure you plan to manage your rights efficiently and beneficially. It is important to include information on copyright status in the relevant metadata, and it is useful for the institution to have a rights management policy that includes guidelines on constructing a copyright statement to be displayed with the digitized work. Apart from the legal reasons for attaching a copyright notice, outlined in Section IV on Rights Management, a clear statement can deter users from misusing the material and, it can be argued, creates an implied contract between the user and the content provider to work within the confines of the statement.

In the desire to protect intellectual property, institutions and copyright holders alike often resort to methods which either compromise user access, or challenge the current concepts of fair use. Although a lot of research and work has been invested worldwide in technologies for protecting intellectual property in the electronic environment (also outlined in Section IV), many institutions still protect their intellectual property by making only low resolution images available on the Internet. Since these are not of sufficient quality to be used in commercial professional printed products, they are less likely to be worth misappropriating. However, although they may be adequate for various online purposes, they cannot support the kinds of detailed research for which high-resolution images are so valuable. Similarly, many owners of copyright materials feel threatened by the possibility of uncontrolled distribution of digital information, and are challenging the concept of 'fair use' in the digital environment, where it is more difficult to control who has access to the material and for what purpose. Fair use lies at the heart of the work of cultural and educational institutions, and many are striving to maintain it in at least its current form. However, you should note that generally, fair use is a safer option within a secured environment (a password protected class website, for example).

An institution must give serious consideration to how much control of its digital assets it is willing to give up or pass to others and examine how the costs involved in licensing and clearing rights would affect the overall budget.

Cultural institutions are exploring other mechanisms for managing the distribution of their intellectual property. Although none of these mechanisms offers a single uniform solution for the diversity of the cultural sector, some useful models are emerging. Licensing is the most popular option at the moment, but many institutions choose either to administer their materials directly, or to arrange for this to be done through some other means, whether through an external agency or through a consortium of rights holders.

(Zorich 1999). Some of these options have existed for a long time, but are relatively new to cultural heritage organizations. The increasing demand for digital copies of the institutions' cultural assets, and the much wider distribution networks that the electronic world has brought, require new strategies from the cultural sector and careful examination of the different options. An institution must give serious consideration to how much control of its digital assets it is willing to give up or pass to others and examine how the costs involved in licensing and clearing rights would affect the overall budget.

Relation with other digitization efforts

With the spread of digitization activity and the multitude of projects and programs around the world, it is important to ensure that your proposed digitization efforts will not duplicate other efforts elsewhere. When selecting material for digitization you should ask yourself if there are any similar or complementary projects in the same institution, country, or even further afield, to avoid a costly waste of effort and resources. By collaborating and coordinating digitization programs, cultural institutions can ensure that they build a critical mass of digital collections. The community needs authoritative, well-governed registries for all types of cultural heritage materials—texts, images, audio collections, etc.—and easy ways to use the registries to determine whether items under consideration have already been digitized. Although such a registry already exists for preservation quality microfilm, there is still no similar listing of digital resources, but a number of initiatives in this area indicate that the situation is likely to change in the future.

Link Box:

Register of Digitization Initiatives and Programs

With the recognition of the need to eliminate redundancy of effort, to maximize the available knowledge about digitization projects and resources, and to improve communication between digitization activities, a number of initiatives in the US are providing mechanisms for projects to register details of their activities and outputs. Among those that have particular value for the humanities are:

- *NINCH* International Database of Digital Humanities Projects http://www.ninch.org/programs/data/
- *ARL* (Association of Research Libraries, Washington DC) Digital Initiatives Database (in collaboration with the University of Illinois - Chicago) http://www.arl.org/did/
- *DLF* (Digital Library Federation) Registry of Digital Reproductions of Paper-based books and serials http://www.diglib.org/collections/reg/reg.htm
- *CETH* (Rutgers University Center for Electronic Texts in the Humanities) Directory of Electronic Text Centers http://harvest.rutgers.edu/ceth/etext_directory/
- *IMAGELIB* (University of Arizona Library) The Clearinghouse of Image Databases. http://www.library.arizona.edu/images/image_projects.html

Until initiatives in this area expand and such a source of information is created, cultural heritage professionals need to use traditional professional and research channels to collect information about existing digitization efforts and future plans. Web searches and relevant portals, word of mouth from colleagues, special-interest professional email lists, related professional and academic conferences and journals (e.g. *DLib* or the *RLG DigiNews*) can provide a wealth of information and a good starting point to map activities in the area.

Link Box:

Web portals on humanities computing resources

- Humanist discussion group http://www.princeton.edu/~mccarty/humanist/
- Humbul Humanities Hub (UK gateway of humanities resources) http://www.humbul.ac.uk/
- Labyrinth (Resources for Medieval Studies) http://www.georgetown.edu/labyrinth/labyrinth-home.html
- Voice of the Shuttle, (the portal on humanities resources maintained by Alan Liu, University of California, Santa Barbara) http://vos.ucsb.edu/
- BUBL: Internet resources covering all academic subject areas, based in the UK. http://bubl.ac.uk/link/

When carefully thought out and planned, collaboration with other institutions (covered in greater depth in Section IX) can be part of an effective strategy for digitization, enabling the sharing of resources and expertise. In fact, digitization activities have often encouraged and led to collaboration with others, as cultural institutions found that this was a way that they could afford expensive specialized equipment, or take advantage of the expertise of computing science departments to create advanced tools for managing digital collections, or enhance their digital collection by collaborating with institutions with complementary analog sources. By taking care that the materials you select for digitization complement initiatives at a local, regional, national, or international level, you can increase the breadth and impact of your work, making a valuable contribution towards truly useful digital resources and tools that answer real needs and are widely used. To this end, we need national or even, in the case of the European Union, trans-national strategies for the digitization of cultural material, to coordinate activities and ensure that we invest in creating unique, complementary, interoperable and high quality digital assets (Ross & Economou 1998).

Metadata and interoperability

Another criterion for selection of material for digitization is the availability and quality of the related metadata. Digitization activities often reveal backlogs and documentation gaps or inconsistencies in the management of the analog collections and the related analog documentation (which is also a form of metadata, even if not in digital form). It is important that you have good knowledge of the state of documentation, cataloging, and metadata used for the analog materials in order to make informed decisions about selection and to establish realistic projections of project costs and timelines. You may want to give priority to materials that have extensive, high-quality documentation and analog metadata in order to carry out the digitization activities without further delays, adding a minimum amount of digital metadata. On the other hand, digitization can provide the impetus for tackling cataloging backlogs or problems that hinder access to parts of the collection. In any case, a good assessment of the situation is important and should be undertaken before the selection stage, in order to allow you to plan both the cataloging and the digitization process accurately.

It is important that you have good knowledge of the state of documentation, cataloging, and metadata used for the analog materials in order to make informed decisions about selection and to establish realistic projections of project costs and timelines.

Another issue that influences selection is the existence of metadata about the digital surrogate. By coordinating not only with existing metadata activities within your organization, but also with national and international initiatives, you can avoid duplicating other efforts, and will be able to take advantage of the extensive work that has been taking place in this area. Have any other departments or staff in your organization

carried out digitization projects before? Did they devise any subject categories or a metadata scheme that you could use and adjust? You need to decide on the type and depth of information you will record about the digital surrogate and how this will relate to the information on the original source. This has staff and resource implications and needs to be taken into account in the program planning stage. Selecting a uniform collection with materials of the same format and type, for example, will require less investment in metadata planning and recording.

There are three main categories of metadata: descriptive, administrative, and structural. You need all three to manage your digital resources. They are discussed in greater detail in the appendix on metadata, but some discussion will be useful here. Descriptive metadata is the information you will need to record in order to identify the digital resource, its analog original if not born digital, or any analog or digital derivatives. Administrative metadata is essential to the management of the digital asset, and describes its creation, intellectual property status, provenance, and the like: for instance, detailed information about technical specifications and the digitization process which can ease future migration paths and ensure consistent results over time. Structural metadata describes the structure of the digital object and the relationships between its components; this information is crucial to assist navigation and ensure that complex objects that belong to a larger collection are linked meaningfully together. The recently developed METS (Metadata Encoding and Transmission Standard) encoding scheme, maintained at the Library of Congress, provides a robust way to record all three types of metadata, and allows the open use of multiple metadata standards simultaneously.

These metadata categories are vital for the longevity and preservation of the digital material, and in applying them you should work not only towards local consistency and completeness, but should also taking into account the work being done on metadata creation at a national and international level. Furthermore, if your digitized collections are to be retrieved easily by search engines and used together with complementary collections, consistent application of widely accepted standards in metadata recording is essential. Although there is no general consensus on the most appropriate metadata scheme for all metadata categories, it is important to keep abreast of the intensive activities in this area as generally accepted models start to emerge (OCLC/RLG 2001). The Dublin Core metadata initiative (http://dublincore.org/), for example, has gained international recognition for its efforts on descriptive metadata to ensure resource discovery and accurate identification in the electronic environment. It also demonstrated the importance, as well as the difficulties, of building general cross-sectoral consensus on metadata issues, which will be necessary in order to create resources that are interoperable across different user groups, disciplines, and institutional types.

(More information on metadata is available throughout the Guide, e.g. in Section V on Digitizing Texts, Section VI on Images, Section VIII on Quality Assurance, Section XIV on Preservation, and in the appendix on metadata.)

Digitization issues

Having examined all the areas discussed so far—the analog objects, their intellectual and physical characteristics, rights management questions, the reasons why these objects would be digitized, and users' needs—you should now assess the issues that all these raise for digitization.

In this process you should determine what features would have to be retained in the digital surrogate. Some of the questions you might ask are listed in the Question Box below. You should also examine those features of the original objects that might cause problems during digitization, such as the objects' size, their physical material, and their state of preservation.

Question Box:

What Features of the Original Should Be Retained in the Digital Surrogate?

- Should the digital copy retain color information?
- Do the images, sound, or video files you will deliver need to be of high resolution?
- Should digitization preserve the background 'noise' in the recording where this results from a live performance?
- Should digitization preserve the background 'noise' in the analog recording where this is a function of the age of the media or the process of production?
- Is it necessary to produce machine-readable text?

Examples of original features that may require special accommodation:

- Oversize maps will require special hardware or software for scanning
- Bound materials will need to be disbound before scanning, if using a sheet-fed scanner.
- Fragile items might require conservation treatment prior to digitization.

What are the technical and resource implications for digitization in order to retain these features or address these problems so that the results are of adequate quality to meet the aims of the program?

- You will need to calculate the file sizes that will be created from scanning in color and at high resolution.
- You will need to consider whether altering or destroying the original material is an acceptable option.
- You will need to assess the time and staff required for disbinding the bound books, and the availability and cost of large format scanners.
- You will need to assess the time, staff, and tools required to create descriptive, structural and administrative metadata.

When selecting a digitization approach, you should always start from your sources and users and aim to match these with the most appropriate procedure. Thus for example, it may not necessarily be appropriate to scan at the highest possible resolution or use a certain piece of equipment simply because it is available. If the technology you need is not accessible or affordable, you may wish to explore the possibility of collaborating with another institution that could offer access to the equipment you need. Alternative strategies, such as digitization from intermediaries, may also be worth considering.

You also need to examine the long-term use and development of the collection and estimate how it will grow. This should already be part of the selection process at an early stage, as future development and scope may influence your selection priorities and digitization strategies. If you know from the start that you will be dealing with a very large collection and will need to accommodate subject retrieval, for instance, you can

build in provision for robust keywording even though this may seem like overkill for your first small collections. Although you might be starting small in the beginning, you will have prepared the ground for future expansion and growth.

Cost-benefit analysis

Another important consideration to take into account is the relationship of projected costs to expected benefits. Institutions in the cultural and educational sectors traditionally operate with limited resources and try to satisfy many competing demands. Although cost-benefit analysis is not the only consideration, it is nevertheless an important one that cultural heritage institutions cannot afford to ignore. However, what we should not forget in the analysis is that benefits from digitization might be intangible, especially when we are dealing with programs in the educational and cultural sector.

As digitization is very resource-intensive, at both the creation and the maintenance stage, you should first examine whether there are lower cost alternatives to digitization for achieving your stated goals. Sometimes, for example, the publication of a traditional scholarly catalog might serve your specialized research community better, or traditional access might be sufficient for resources that are of mainly local interest to a small group.

Even if you do decide to digitize, there might be lower cost alternatives to the digitization approach you have selected. For example, is color scanning really necessary to serve your aims and users? This is the most expensive approach (creating the largest files), compared to grayscale (which records shades of gray using 8 bits per pixel, as explained in Section VI on Images) or bitonal scanning (which records information only in black or white using only one bit per pixel). Bitonal scanning, which creates the smallest file sizes and is the generally most affordable approach, might be sufficient for scanning text, although there are documented cases of unbound grayscale scanning with a flatbed or sheetfeed scanner being less expensive per page than bound bitonal scanning with overhead scanners or digital cameras. Similarly, in some cases, uncorrected OCR text might be sufficient for retrieval and indexing. Hand-correcting the output of OCR software improves retrieval accuracy, but significantly increases the costs of text conversion. It may even be worth asking whether you need to perform OCR in the first place, or whether it might be sufficient to provide an index with links to the page images.

There are a number of challenges and even uncertainties in conducting a careful cost-benefit analysis. For instance, the labor costs associated with scanning are usually much lower than those related to selecting, preparing, inspecting, and indexing digital resources. Indeed, it is generally true that the intellectual overhead for digitization efforts—the cost of setting up infrastructure, making decisions, overseeing quality control, ensuring good quality metadata—is among the most significant costs. For this reason, it has been suggested that it is more economical to convert once at a high level in order to avoid the expense of duplicating the process at a later stage when more advanced technology requires better quality digital objects (Lesk 1990). However, these claims have not yet been borne out clearly in practice. Similarly, it can be extremely difficult to calculate expenses and compare the anticipated costs of new projects with those of

existing projects. Again, the main costs are usually associated with the staff involved in digitization but this can vary enormously.

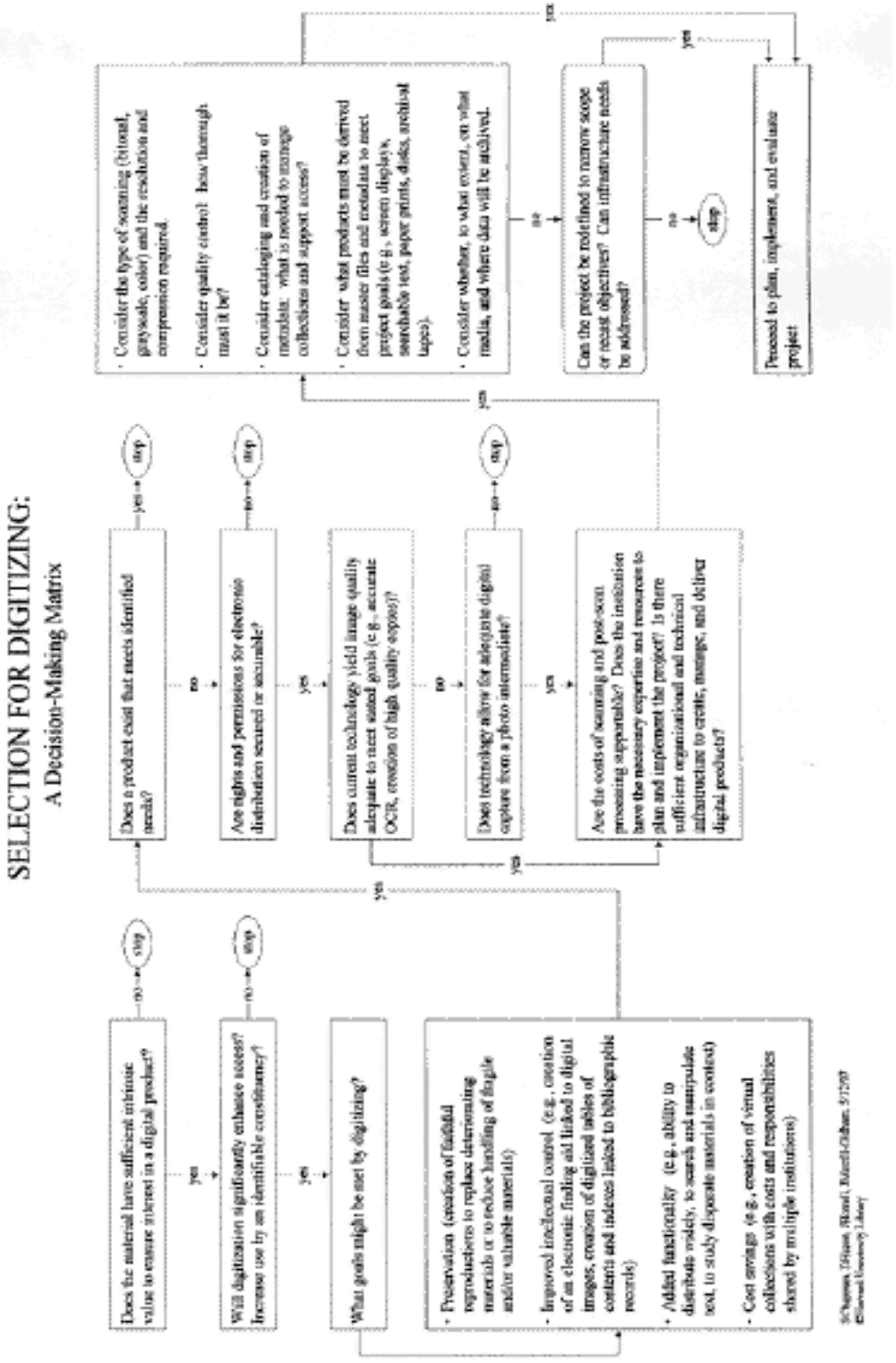
There also are many hidden costs that are often omitted in published reports. Cataloging, indexing, preparation of material for scanning, post-scanning processing of the material, quality assurance, and maintenance of digital resources are some of the activities that are not always calculated or indicated separately. “Though digitizing projects must calculate the likely costs and benefits, our ability to predict either of them is as yet rudimentary” (Hazen et al 1998). Another area of hidden digitization costs, ironically, is collaboration. Although, as we mentioned above, working with other institutions can potentially save money through the sharing of expensive equipment, it can also involve added commitments of time and resources—for instance, the overhead required in coordinating joint efforts, attending meetings, and the like. If not planned carefully, these may be missed in the project planning and create unforeseen cost overruns. Finally, in the calculation of costs and benefits other important factors are the levels of usage and the distribution mechanisms selected (the latter are discussed in greater depth in Section X on Distribution).

Digitization might lead to some cost savings, though these are unlikely to offset the project costs fully. For example, digitization may reduce the need for storage space, as is the case with the JSTOR project (http://www.jstor.org). This project helps the participating academic libraries reduce the costs associated with the storage and care of journal collections by digitizing and providing easier access to backfiles of over one hundred and fifty journals. Of course, this is no solution for museums where holdings are unique and cannot be discarded after digitization. Cost savings might also be involved in reducing staff time spent in retrieving information in analog form, but on the other hand, digitization creates a whole series of new demands and costs, particularly for the management and long-term preservation of the digital resources.

If digitization is indeed the best option, but involves higher costs than you can afford, you should examine whether you can secure external funding (see also Section XI on Sustainability). The priorities of funding bodies have often influenced the selection of material for digitization, with grant applicants trying to find out what is more likely to receive funding and selecting projects and material with that in mind. This is also a consideration with sponsors and donors of material, who often have their own agenda and can influence selection priorities. In some cases, however, you might find that the areas supported by external funding organizations agree with your institutional goals. Having a strategic knowledge of the collections and development plans, as discussed at the beginning of this section, can help you take a constructive approach in seeking sponsors and donors, so that their priorities match those of your institution.

The issues discussed here represent some of the questions you would need to examine with your project team when defining the selection criteria for your program. The criteria you identify will depend on the particular characteristics of your institution and the aims of the program. Once you have agreed on some key principles and guidelines, it is very useful to document them and share them among the team members. The diagram on the

next page, prepared by Harvard University Libraries, is one possible model that summarizes some of the questions and decisions involved in selecting material for digitization.


SELECTION FOR DIGITIZING:
A Decision-Making Matrix
Does the material have sufficient intrinsic value to ensure interest in a digital product?
Will digitization significantly enhance access? Increase use by an identifiable constituency?
What goals might be met by digitizing?
• Preservation (creation of faithful reproductions to replace deteriorating materials or to reduce handling of fragile and/or valuable materials)
• Improved intellectual control (e.g., creation of an electronic finding aid linked to digital images, creation of digitized tables of contents and indexes linked to bibliographic records)
• Added functionality (e.g., ability to distribute widely, to search and manipulate text, to study disparate materials in context)
• Cost savings (e.g., creation of virtual collections with costs and responsibilities shared by multiple institutions)
Does a product exist that meets identified needs?
Are rights and permissions for electronic distribution secured or securable?
Does current technology yield image quality adequate to meet stated goals (e.g., accurate OCR, creation of high quality enlarges)?
Does technology allow for adequate digital capture from a photo intermediate?
Are the costs of scanning and post-scan processing supportable? Does the institution have the necessary expertise and resources to plan and implement the project? Is there sufficient organizational and technical infrastructure to create, manage, and deliver digital products?
• Consider the type of scanning (bitonal, grayscale, color) and the resolution and compression required.
• Consider quality control: how thorough must it be?
• Consider cataloging and creation of metadata: what is needed to manage collections and support access?
• Consider what products must be derived from master files and metadata to meet project goals (e.g., screen displays, searchable text, paper prints, disks, archival tapes).
• Consider whether, to what extent, on what media, and where data will be archived.
Can the project be redefined to narrow scope or recast objectives? Can infrastructure needs be addressed?
Proceed to plan, implement, and evaluate project
skip
yes
no

http://www. clir.org/pubs/reports/hazen/matrix.html or
http://preserve.harvard.edu/bibliographies/matrix.pdf

SELECTION FOR DIGITIZING: A Decision-Making Matrix from Dan Hazen, Jeffrey Horrell, Jan Merrill-Oldham, Selecting Research Collections for Digitization, Council on Library and Information Resources (August 1998), http://www.clir.org/pubs/reports/hazen/pub74.html

---

[1] The American Memory Project received over 40 million requests per month in 1999, increasing to over 50 million in 2000, and over 80 million in 2001, numbers that exceed by far the number of readers who visit the reading rooms of the Library of Congress.

IV. Rights Management

> Libraries want to share content; publishers want to sell it. Museums strive to preserve culture, and artists to create it. Musicians compose and perform, but must license and collect. Users want access, regardless of where or how content is held. What all of these stakeholders (and more) share is the need to identify content and its owner, to agree on the terms and conditions of its use and reuse, and to be able to share this information in reliable ways that make it easier to find. (Bearman et. al., 1999)

Introduction

Intellectual property rights (IPR) include copyright, trademarks, patents, publicity rights, privacy, and trade secrets, but it is copyright that will mostly concern this audience. Cultural institutions are primarily interested in two of the many issues that surround copyright: how they can legally digitize material in which they may not hold the copyright and how they can ensure that no one else can use the materials they have digitized without their approval (tacit or otherwise).

Copyright

Broadly speaking, copyright grants exclusive but limited rights to the creator of an original work to copy, reproduce, perform and distribute it. From its inception, however, copyright law has been as much about the promotion and circulation of knowledge and good ideas as it is about the protection and rewarding of creators. Limitations and exemptions to creators' copyright protection are as important to society as the protection itself. Certainly in the United States, the balance between the rights of creators and the rights of society in general is of key importance.

Categories of material

In the United States, "original works of authorship" protected by copyright include:

- literary works, which covers all kinds of literature (software is included in this category);
- dramatic works, including any accompanying music;
- musical works, including any accompanying words;

- artistic works;
- pantomimes and choreographic works;
- pictorial, graphic, and sculptural works;
- motion pictures and other audiovisual works, whatever their medium (film, television, DVD, etc.);
- sound recordings, which covers the spoken word;
- architectural works.

Practice and case law have tended to view these categories in their broadest sense. Maps are protected as 'pictorial, graphic, and sculptural works', and computer programs are protected as 'literary works'. This categorization may affect certain rights; for instance, ideas, facts, short phrases, and blank forms are excluded from copyright protection, which focuses on the particular expression of ideas in a tangible medium.

Copyright and Digitization

Before carrying out any digitization, institutions need to establish the copyright status of the source material that will be digitized. If this investigation shows that the institution does not itself hold the copyright in the material, then the institution has three options: (a) abandon plans to digitize the material, (b) secure permission to digitize the material; or, (c) proceed with the project anyway under one of the exemptions to the exclusive rights of the copyright owner found in U.S. copyright law, such as the fair use exemption, but on the understanding that this involves an assessment of the risks of doing so (see Managing IPR Risk section, below). Projects using derivatives, such as photographic reproductions, rather than originals for digitization, need to examine the copyright status of both the derivative and the original. They may find, for instance, that by digitizing a photograph of a work of art they are infringing the rights of both the photographer and the artist who created the original work.

Once the project team establishes the rights status of the original and derivatives they plan to digitize, they should declare this clearly in the metadata associated with the digital resource and/or the relevant web page.

When considering copyright ownership be particularly aware that there are layers of distinct rights in material in which a photographic reproduction, rather than an original, is being used to make a digital surrogate. The copyright status of both the original and the photographic derivative need to be determined. An excellent guide to negotiating the layers of rights related to photographs of visual materials is the "Copy Photography Computator" (http://vraweb.org/computator/welcome.html) produced by the Visual Resources Association.

Duration of protection and public domain

Any digitization project should begin with an analysis of who owns the copyright in the work to be digitized. If your evaluation of the copyright status of the material you hope to digitize reveals that your institution does not own the copyright, it may still be possible to digitize it if the rights in the material have passed into the public domain. For most categories of material, and particularly for literary and artistic works created after 1977, copyright protection generally lasts for *seventy years* after the death of the author/creator. When this term expires, the work enters the public domain where "all entities, information and creative works are available for use by anyone for any reason without restriction" (Zorich 2000). Note that there can be moral rights issues that exist independently of copyright term under certain foreign laws (see discussion of moral rights below). To avoid moral rights problems, avoid drastic cropping of visual works or other alterations and avoid removing the names of artists and authors of original work you reproduce.

The following chart prepared by Laura Gasaway, Director of the Law Library and Professor of Law at the University of North Carolina, summarizes the relevant terms and duration of protection according to the U.S copyright law, depending on the date when the material was produced or published. The chart makes clear the implications of the new copyright term that came into effect in the U.S. with the 1998 Copyright Term Extension Act. By extending U.S. term from 50 years after the death of the author to 70 years, and altering the special terms for certain published and unpublished works, this Act effectively put a twenty-year moratorium on works entering the public domain (Zorich 2000). For example, works published in 1923 that were next in line to enter into the public domain have been delayed until January 2019 as a result of the Act. Unpublished works (for example, historical documents like letters, diaries and manuscripts) created before 1978 will lose copyright protection in January 2003 unless they are published before that date. Previously unpublished works published between 1978 and 2003 are granted extended protection to December 2047.

WHEN WORKS PASS INTO THE PUBLIC DOMAIN

Includes material from U.S. Term Extension Act, PL 105-298

DATE OF WORK	*PROTECTED FROM*	*TERM*
Created 1-1-78 or after	When work is fixed in tangible medium of expression	Life + 70 years[1] (or if work of corporate authorship, the shorter of 95 years from publication, or 120 years from creation[2])
Published before 1923	In public domain	None
Published from 1923 - 63	When published with notice[3]	28 years + could be renewed for 47 years, now extended by 20 years for a total renewal of 67 years. If not so renewed, now in public domain
Published from 1964 - 77	When published with notice	28 years for first term; now automatic extension of 67 years for second term
Created before 1-1-78 but not published	1-1-78, the effective date of the 1976 Act which eliminated common law copyright	Life + 70 years or 12-31-2002, whichever is greater
Created before 1-1-78, but published between then and 12-31-2002	1-1-78, the effective date of the 1976 Act which eliminated common law copyright	Life + 70 years or 12-31-2047 whichever is greater

[1] Term of joint works is measured by life of the longest-lived author.

[2] Works for hire, anonymous and pseudonymous works also have this term. 17 U.S.C. § 302(c).

[3] Under the 1909 Act, works published without notice went into the public domain upon publication. Works published without notice between 1-1-78 and 3-1-89, effective date of the Berne Convention Implementation Act, retained copyright only if, e.g., registration was made within five years. 17 U.S.C. § 405.

Notes courtesy of Professor Tom Field, Franklin Pierce Law Center

Chart prepared by Laura Gasaway, University of North Carolina,
Last updated 9-18-01, http://www.unc.edu/~unclng/public-d.htm

Work for hire

Although in some countries only an individual can be recognized as the author of a work, in some countries, such as in the U.S. and the UK, the author can also be an organization. Unless otherwise specified by contract, copyright in any material (whether literary, graphical or photographic) created by staff at universities, as part of their normal duties will belong automatically to the university as the employer. In addition to copyrights created by employees, employers can also own copyrights created by independent contractors when the work falls into one of the narrow categories of work defined by the copyright law as a "work for hire" and when there is a written agreement between the parties that the work will constitute a "work for hire." It is likely, therefore, that if your institution were to commission a photographer to take a photograph of an original artwork, the copyright in the photograph would reside with the photographer unless specific actions are taken to transfer the copyright to the institution. For this reason, it is important to clarify these issues from the outset and include appropriate transfers of copyright in all relevant contracts. For works commissioned before digitization became a common practice, the current rights position must be considered, and in some cases re-negotiation may be necessary. If you are outsourcing digitization work, make sure that the contract specifies that you hold any rights resulting from the digital files created by the contractor.

Fair use

Cultural institutions may wish to digitize materials that are not in the public domain, whose copyright they do not own. In this case, they should examine whether the material and the way they plan to use it may be covered by 'fair use'. 'Fair use' is an exemption under U.S. copyright law that allows one to legally use copyrighted material without the explicit permission of the copyright owner. Fair use is named differently by different national legislations. Different nations also vary on what can be covered by fair use, but these usually include non-profit educational use and private research and study. In the U.S., fair use is one of a set of exceptions to exclusive rights and is framed by four key factors:

1. the purpose and character of the use (e.g. whether such use is of a commercial nature or is for not-for-profit educational purposes; whether the use is transformative);

2. the nature of the copyrighted work (e.g. whether it is based on facts or is an imaginative work);

3. the amount and substantiality of the portion used in relation to the copyrighted work as a whole (a portion of a book rather than the whole, for example);

4. the effect of the use upon the potential market for the copyrighted work.[1]

In an effort to address the interpretation of these four factors, the Clinton Administration established the Conference on Fair Use (CONFU) in 1994. CONFU brought together intellectual property owners and users to negotiate new guidelines for the fair use of digital material in nonprofit, educational contexts. The Conference set up several working groups to investigate the issues, but when CONFU concluded in 1998, none of the groups had produced guidelines acceptable to all parties. Most major educational and cultural organizations were critical of the draft guidelines that had been prepared and opted not to endorse them. In some cases, it was felt that the Guidelines asked the right questions, but provided the wrong answers.

Some institutions decided to review and adjust the Guidelines or design new ones of their own to meet their needs and fulfill a wider strategy of IPR management. One example is the Visual Resource Association's *Image Collection Guidelines: The Acquisition and Use of Images in Non-Profit Educational Visual Resources Collections*, a guide that was developed as a result of VRA's involvement in preparing the CONFU Digital Images Guidelines. Another example, adapted from the CONFU guidelines, are the short and clear "Rules of Thumb" for interpreting the fair use of copyrighted materials devised by Georgia Harper for the University of Texas System. Finally, a third example of an alternative to the CONFU Guidelines is the following "Checklist for Fair Use," prepared by the Indiana University Copyright Management Center (Indiana University-Purdue University Indianapolis).

Link Box:

Alternatives to the CONFU Guidelines

Visual Resource Association, Image Collection Guidelines: The Acquisition and Use of Images in Non-Profit Educational Visual Resources Collections: http://www.vraweb.org/copyright/guidelines.html

Georgia Harper, "Rules of Thumb", adapted from the CONFU Guidelines: http://www.utsystem.edu/ogc/IntellectualProperty/roftimag.htm

Indiana University Copyright Management Center, "Checklist for Fair Use": http://www.iupui.edu/~copyinfo/fuchecklist.html

Checklist for Fair Use

Please complete and retain a copy of this form
in connection with each possible "fair use" of a copyrighted work for your project.

Name: _____ Date: _____

Institution: _____ Project: _____

PURPOSE

Favoring Fair Use	*Opposing Fair Use*
Teaching (including multiple copies for classroom use)	Commercial activity
Research	Profiting from the use
Scholarship	Entertainment
Nonprofit Educational institution	Bad-faith behavior
Criticism	Denying credit to original author
Comment	
News reporting	
Transformative or Productive use (changes the work for new utility)	
Restricted access (to students or other appropriate group)	
Parody	

NATURE

Favoring Fair Use	*Opposing Fair Use*
Published work	Unpublished work
Factual or nonfiction based	Highly creative work (art, music, novels, films, plays)
Important to favored educational objectives	Fiction

AMOUNT

Favoring fair use	*Opposing fair use*
Small quantity	Large portion or whole work used
Portion used is not central or significant to entire work	Portion used is central to work or "heart of the work"
Amount is appropriate for favored educational purpose	

EFFECT	
Favoring Fair Use	*Opposing Fair Use*
User owns lawfully acquired or purchased copy of original work	Could replace sale of copyrighted work
One or few copies made	Significantly impairs market or potential market for copyrighted work or derivative
No significant effect on the market or potential market for copyrighted work	Reasonably available licensing mechanism for use of the copyrighted work
No similar product marketed by the copyright holder	Affordable permission available for using work
Lack of licensing mechanism	Numerous copies made
	You made it accessible on Web or in other public forum
	Repeated or long term use

Prepared as a service of the Copyright Management Center at Indiana University located on the campus of IUPUI http://www.iupui.edu/~copyinfo/ rev. 3.99

Created: 26 July 1999, ARK **Last Updated:** 25 January 2001, LDB

Comments: copyinfo@iupui.edu URL: http://www.iupui.edu/~copyinfo/fuchecklist.html

Institutional guidelines can help project managers make consistent decisions on what can be considered ‘fair use’ in their everyday work. However, it is important to remember that it is ultimately a court of law that determines whether a particular use, when challenged, is ‘fair use’ or not, according to the particular circumstances of each case. Some uses which may seem as if they are clearly ‘fair use’ might be challenged.

Link Box:

For more information on 'Fair Use', good starting points are:

- Stanford University Libraries: Copyright & Fair Use page: http://fairuse.stanford.edu
- Duke University, Center for Instructional Technology: Copyright Law and Fair Use: http://cit.duke.edu/resource-guides/development-copyright.html
- The CONFU Final Report by Bruce A. Lehman, November 1998 is available on the U.S. Patent and Trademark Office website: http://www.uspto.gov/web/offices/dcom/olia/confu/

Moral rights

Even when a cultural institution or publisher owns the copyright in a piece of literature or art, in many countries the creator (e.g. the artist or author) retains the moral right (*droit moral*) to be identified as the creator of the work and is granted protection against derogatory treatment of his or her work. For cultural institutions this means, for example, that it is essential to ensure that the original artist's name be linked with his or her creation (and the use of appropriate metadata is very useful in this direction), that works be normally used in their entirety, and that they not be amended (e.g. digital copies should not be cropped or edited). These issues should be covered by institutional policy in clear guidelines for use. In the U.S., moral rights are limited to the 1990 Visual Artists Rights Act, which recognizes visual art authors' right of attribution, their right of integrity, and their right to prevent the destruction of copies of the work[2]. However, not all moral rights fall within copyright laws as in the U.S., and not all moral rights terms are finite. In some instances, as in France, moral rights may survive the expiration of copyright. Thus one must be careful not to alter or denigrate a reproduction beyond recognition or in a context that may be objectionable to the artist or artist's estate.

Seeking permission to use copyrighted material

In order to legally use material that is not in the public domain and is not covered by 'fair use' or otherwise exempt from liability for infringement, you must acquire permission from the rights holder(s) (unless the material has been licensed for your use and the permissions have been assigned as a part of the license). Depending on the nature of the work or the country of origin of the author or artist, there are numerous collective rights organizations that can help you to clear these rights. The Copyright Clearance Center (CCC) (http://www.copyright.com) for parts of books or journals is one such example.

Link Box:

The Getting Permission page by Georgia Harper at the Office of General Counsel, University of Texas System (http://www.utsystem.edu/OGC/IntellectualProperty/PERMISSN.HTM) includes links to collective rights organizations you can contact to gain permission about the following types of materials:

- Images
- Journalism
- Music Performances
- Plays
- TV News Programs
- Movies

It also includes links to resources for tracing the rights owners of archival materials, such as historical photographs, architectural drawings, or personal papers.

When you do not know who the rights owners are and colleagues cannot offer any additional information, you can contact the U.S. Copyright Office (http://www.loc.gov/copyright/), which provides online searching of some of its records on registration and copyright ownership documents. It also carries out professional searches for a fee. Once you know who the rights owner is, you can contact them directly to ask for permission to use the material. Apart from your contact details, you should describe clearly:

- the work you want to use (with a copy if possible);
- the scope of the project;
- where and how you intend to use the work (e.g. names of key contributors, approximate size of the project, URL, anticipated life of the project, how it is going to be distributed);
- any future use you envision (e.g. if you are preparing a CD-ROM for educational use, you might want to consider web delivery too);
- the specific rights you seek (e.g., for use in all classes you teach, or your department teaches; for all editions and future revisions of the published CD-ROM; for USA, North American, or world rights; for English-language specific languages, or all languages; for print, performance, display, or electronic rights).

You should ask for instruction on the wording of the credit line, the copyright notice related to their material, any other conditions they might have, and the fees that might apply. You should also ask for confirmation that they have authority to grant permission and if not, you should request that they direct you to the appropriate rights-holders. You may request them to warrant that your use will not, to their knowledge, infringe the rights of any other third party.

Make sure to document all your efforts to trace the rights-holders, since if they prove to be untraceable or unresponsive and you decide to go ahead with the project, such documentation could help to prove “good faith best efforts,” or “due diligence” if original rights-holders were to initiate legal proceedings at a later date. Such evidence might enable you to argue that the economic impact was limited as the owner(s) made it difficult to license the information.

Licensing schemes

Given the considerable effort and resources that cultural organizations expend in creating digital assets, issues of information control, licensing structures, and royalties deserve careful attention. Strict security measures can constrain and frustrate legitimate users, while rights management tools currently require a significant investment. Additionally, even the best technology-enabled protection schemes can be defeated by persistent

hackers. And, while such hacking is illegal under the DMCA, pursuing infringers in court is costly and time consuming.

After considering these issues, organizations normally conclude that it is more sensible to develop an effective business model for the use of digital collections. Many of these business models depend upon licensing, which can be helpful either in obtaining permission to digitize material or in distributing it to others. Site licensing is one model for managing rights in and access to digital resources. Organizations have been experimenting with different licensing schemes for several decades. These schemes usually take into account the number of users and the types of institutions, for example by offering special educational licenses to universities. Licensing provides one fairly secure way of protecting rights and ensures that the holders (or their distributor) can track the users. Another effective mechanism may be to join rights-holders' consortia, as was mentioned in the IPR Management section of Section III: Selecting Materials, which also discusses briefly the issues that consortia raise.

Example Box:

Examples of Licensing Schemes

SCRAN: an example of a licensing scheme

SCRAN, the Scottish Cultural Resources Network (http://www.scran.ac.uk), which has supported the digitization of cultural assets by museums, libraries, and archives, has created a large multimedia resource about Scottish culture and history. SCRAN has developed a Contributor License in which the contributor (e.g. a museum, library, or archive) retains all commercialization rights in the new digitized object, but grants to SCRAN a perpetual, non-exclusive right to use the digital object for any non-profit educational purpose, worldwide. SCRAN passes back to the contributor an agreed payment for any commercial use of the digital object. If SCRAN were to be taken over by any commercial third party, all the educational rights would return to the original contributor. With the User License, SCRAN grants a non-transferable, non-exclusive right to use of the digital objects by the employees and students of the user institution and any member of the public visiting the institution for educational purposes. In addition, the project has devised an extensive range of usage fees, which covers individuals, schools, higher education institutions, libraries, museums, and commercial users and takes into account their size and usage. All resources are digitized at a very high resolution. 'From this archival resource, a network surrogate is created at a lower resolution, to minimize download time. This networked resource may only be downloaded by members of an educational institution licensed by SCRAN. It is protected by an invisible "watermark" (to confirm the resource's copyright status) and "fingerprint" (to identify who downloaded it and when). To avoid any accusations of "entrapment", this information is also clearly shown in banners at the top and bottom of the downloaded image. But any member of the public, anywhere in the world, has access to a thumbnail image of the asset, plus full textual documentation' (Royan 2000).

AMICO: collective licensing

The Art Museum Image Consortium (http://www.amico.org) is an independent nonprofit consortium of organizations with collections of works of art that seek to deploy the digital documentation of those works for educational use. Members are collectively building the AMICO Library as a centralized digital educational resource that is licensed, under subscription, to universities and colleges, public libraries, elementary and secondary schools, museums and galleries.

When subscribing to the AMICO Library, educational institutions select a distributor, such as the Research Libraries Group, SCRAN or H.W. Wilson. Authorized users from AMICO subscribers and members then access the Library through a secure electronic distribution system. The public web site does not allow access to the full Library, but includes thumbnail images and brief text descriptions of all works. Educational institutions are charged an annual fee to subscribe to the Library, which provides unlimited access to the institution's authorized users for a year.

International perspective

Although cultural organizations operate in an age of instant, global communications, copyright law is still largely earth-bound, the creature of national legislation. In general, countries establish their own copyright laws and enforce those rules within their territories. For example, Country A might provide for a copyright term of life of the author plus 70 years, while Country B may set the term at life of the author plus 50 years. Within this framework of national laws, a number of international treaties (such as the Berne Convention and the Universal Copyright Convention) establish certain universal "minimum standards," harmonize many of the disparate aspects of national copyright

laws, and ensure that protection is available across national borders on a non-discriminatory basis.

Nonetheless, cultural organizations may encounter complex legal issues when infringement occurs across national borders. Suppose that a cultural organization located in Country B digitizes (without permission of the author) a copyrighted book written by a citizen of Country A. Since both Country A and Country B are members of the Berne Union, the author from Country A will be entitled to the same rights (and subject to the same limitations) as Country B extends to its own authors.

In general, the law of the country where the infringement took place (and where relief is sought) will govern. Thus, to continue the example, if the law of Country B excuses the infringement under a broad fair use provision, the author may be out of luck even though the law of Country A contains no similar defense. Moreover, in an electronic environment, where material is transmitted from one country to another with the press of a button, determining the place of infringement (and applicable law) itself may be difficult. In analyzing situations involving possible transnational infringements, cultural organizations should begin by posing the following questions:

- in what countries does the institution seek protection?
- what protection is available under that country's national intellectual property laws?
- what treaty provisions may provide protection or facilitate obtaining intellectual property rights in the absence of or as a supplement to local laws? (Shapiro & Miller 1999)

Link Box

Developments to Watch:

- the 2001 ‘European Directive on Harmonization of Copyright’ which tries to harmonize European law: http://europa.eu.int/eur-lex/en/dat/2001/l_167/l_16720010622en00100019.pdf. For a commentary on some of its key elements and how it might affect cultural institutions see the NINCH website: http://www.cni.org/Hforums/ninch-announce/2001/0075.html.
- the Hague Conference on Private International Law that aims to address jurisdiction issues: http://www.cptech.org/ecom/jurisdiction/hague.html.
- the 1994 ‘Uruguay Round Agreements Act’, which, among others, restores copyright in foreign works that have entered the public domain in the U.S: http://www.loc.gov/copyright/gatt.html. Diane Zorich (2000) discusses some of the implications of the Act for the cultural sector in an article that can be found at the NINCH website: http://www.ninch.org/copyright/2000/chicagozorich.html.

Managing IPR risk

Protecting and managing intellectual property rights, and avoiding their infringement in the cultural sector, all involve risk management. When cultural institutions design digital collections, they may well select material in which they do not hold copyright. It is easy to imagine a situation where material an institution would like to use has not entered the public domain, the kind of use it would like to make of the material is not covered by 'fair use', and it has not been able to obtain permission to use the material. For example, a local history museum is creating a new CD-ROM for sale and wants to include digital images of a series of landscapes which were discovered in the attic of the local arts college. These were created after 1980, judging by the buildings that some of them portray, by an artist whose identity is not known to the museum. Despite showing the material to college staff and searching relevant records and making best endeavors (all diligently documented) to contact students who attended the college at that time, the museum does not succeed in identifying the artist and tracing the copyright holder(s). In this case, the staff would have to assess the risk of proceeding and balance the benefits to be derived from using these particular materials for the project, against risks such as negative publicity, costs of litigation, financial penalties that might be awarded, costs of lost human resources (e.g., administrative time), or financial loss to be incurred by having to withdraw the digital resource. In some cases, the staff might conclude that there are important considerations favoring limited use of the material that would counterbalance the risk of infringing the legal rights of the unidentified copyright owner.

The discussion of Project Management in the Resources section encourages institutions to produce a risk table as part of good project planning. One large category of risk that would be covered in that table relates to rights issues. Here we repeat the advice in that Section, but focus it more on evaluating issues associated with rights.

- What can go wrong? (e.g., What could be the consequences of using the material without specific permission, pursuing a fair use or mitigation of damages argument?)

- For each risk, what is the likelihood of actual risk for the project? (e.g., Has a thorough search to trace the rights holder been made without any results? How can you demonstrate that this search was conducted in a comprehensive manner? Have these efforts been documented?)

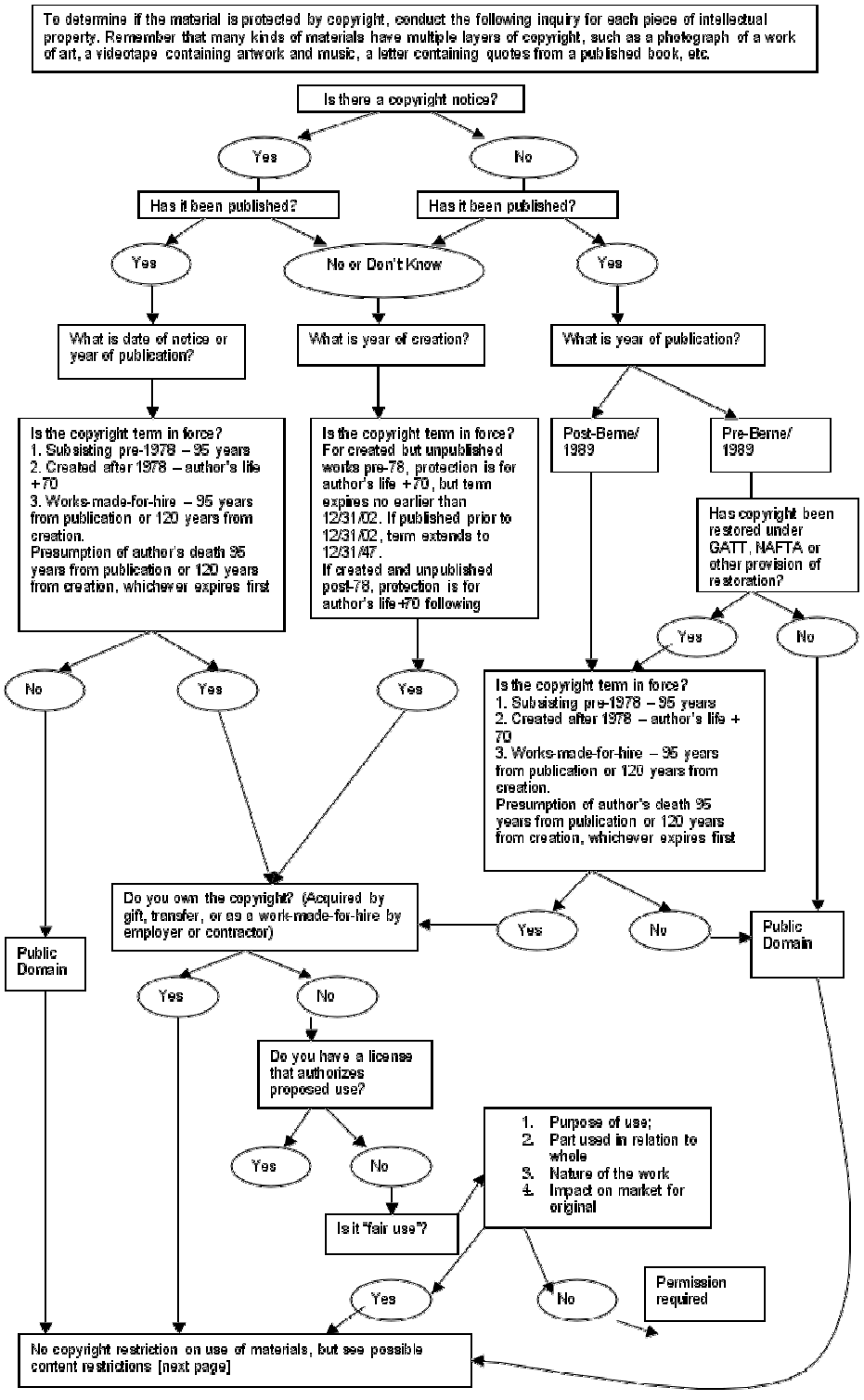
- What can be done to avoid the risk or to minimize its impact? (e.g. Have these efforts been documented?)

- If copyright infringement does occur, what would be the impact on the project/program? (e.g., What would be the impact in terms of scale in time, finance or reputation?)

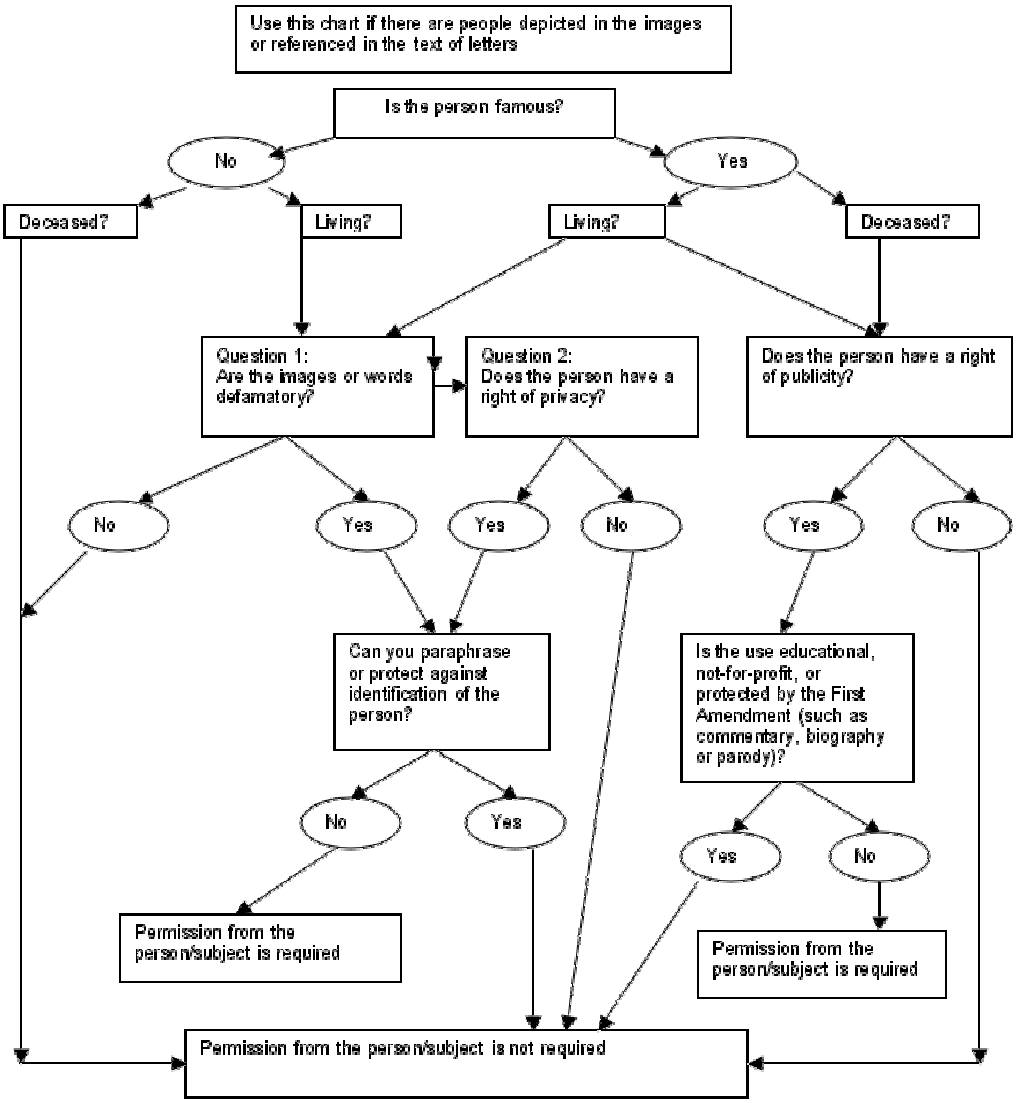
- How will the institution address the risk if it does occur? (e.g. Will it withdraw the material or negotiate with the right holder? Will the institution be able to afford legal expenses?) [3]

As one expert suggests “the test [we] ought probably to apply, is not ‘is this legal?’, but ‘does this harm any one?’. If an action does not harm the financial interests of another institution or individual, it is probably safe to take it” (Royan 1998). While there may be merit in Royan's argument, especially where your institution can demonstrate that it has acted with due diligence to identify the holder of the rights without success (see above), cultural institutions must be aware of copyright regulations and act prudently to avoid explicitly infringing them. So before investing considerable effort in digitizing your collections, you should address the questions outlined in the following decision tree diagram, where we summarize the questions raised throughout this section:

1. REPRODUCTION OF MATERIALS ON WEB



2. CONTENT RESTRICTIONS FOR REPRODUCTION OF MATERIALS ON MUSEUM WEB SITE



Right of Privacy - the right to be left alone, not portrayed in a false light, not have private facts publicly divulged, not have life intruded upon, and not have name or image misappropriated. The right applies only to living persons.

Defamation - the right to prevent false and malicious comments about you from being published.

Right of Publicity - the right of celebrities to control the use of their name, image, likeness, signature and voice. The scope of the right and how long it lasts are subject to state law.

Trademark - cannot use trademarks or logos for a commercial purpose or if the use creates confusion as to the source of the product or service.

Link Box:

Rights Management Resources

There is a wealth of available information on rights management issues on the Web. Good starting points include:

- U.S. Copyright Office, Library of Congress: http://Lcweb.loc.gov/copyright
- American Library Association Washington Office: Copyright & Database protection page: http://www.ala.org/washoff/ip.html
- Shapiro, Michael & Miller, Brett 1999. A Museum Guide to Copyright and Trademark. American Association of Museums. Highlights of the book are available on the AAM site: http://www.aam-us.org/resources/reference_library/mus_guide_copyright.cfm.
- McCord Hoffman, Gretchen. 2001. Copyright in Cyberspace: Questions & Answers for Librarians. Neal-Schuman Publishers.
- The Digital Millennium Copyright Act of 1998 (http://www.loc.gov/copyright/legislation/dmca.pdf) implemented the treaties signed in December 1996 at the World Intellectual Property Organization (WIPO) Geneva conference, but also contains additional provisions addressing related matters (some of which were opposed by many scientists, librarians, and academics). For a discussion of the key points, see the site of the UCLA Online Institute for Cyberspace Law and Policy: http://www.gseis.ucla.edu/iclp/dmca1.htm.
- The Digital Future Coalition (http://www.dfc.org) is a group 'committed to striking an appropriate balance in law and public policy between protecting intellectual property and affording public access to it'. The Coalition brings together non-profit educational and cultural organizations together major commercial associations from the computer, telecommunications, and network access industries.

Rights of privacy

The massive availability of digital resources has led to the exposure of much previously unpublished material, including personal, commercial or sensitive information. In the selection stage, but also when considering public access to digitized material, the project staff should ensure that the material is handled with responsibility, sensitivity, and care. Projects that involve the digitization of personal data should be aware of the data protection legislation of the country where the project is based. The different notions of privacy in different nations' laws will continue to have an uncertain impact in the cultural heritage field. Some general key guidelines are:

- Personal data should be obtained lawfully;
- Personal data should be collected, kept, and used for specified purposes;
- Rights of access and of processing personal data should be clearly specified;

- Where users are provided access to personal data, your institutions should maintain an audit trail to track who has accessed the data and how and when they used it.

In the U.S., privacy on the Internet is largely self-regulated. Increasingly, however, companies are shunning sites and associations that do not follow basic privacy safeguards. The Direct Marketing Association has gone so far as to expel members who do not follow basic privacy safeguards, for instance those who do not post privacy statements. The industry is attempting to avoid federal legislation in this area through self-regulation. Although the Federal Trade Commission (FTC) has so far not recommended any Internet privacy legislation except with respect to children, a provider of information — who collects personal information should disclose the following information: why it is being collected; how it will be used; what mechanisms are in place to protect its confidentiality, ensure its quality, and guarantee its integrity; what impact users' decisions to provide or withhold information will have on the services they can access; and, what redress individuals have if inaccurate information is disclosed. If you are a provider of information and collect any information about your users you should have a privacy policy on your website and it should be clearly labeled and easily accessible to users.

Institutional policies can be very useful for navigating this complex area of information privacy. For example, the information privacy policy of the Science Museum of Minnesota (http://www.sci.mus.mn.us/museum/identity/Privacy.html), covers, among other things, information collected by the Museum and submitted by third parties, personal and other information collected on its website, the use and sharing of information by the Museum (including use of postal addresses, telephone numbers, and emails and how to opt-out), the use of cookies, and children's privacy on the web (with reference to the Children's Online Privacy Protection Act).

Example Box:

In the case of the Survivors of the SHOAH Visual History Foundation (VHF), the copyright in the interviews with the Holocaust survivors has been assigned to the VHF. Each interviewee signed a release that granted the Foundation the rights to distribute the testimonies for educational purposes. VHF recognizes its duty to protect the sensitive material contained in the testimonies from misuse and misrepresentation. This duty is reflected in the approach that the VHF takes to the distribution of the testimonies. The VHF has identified four categories of use:

- ***Foundation-authenticated use,*** by users who are using the archive within the Foundation itself, who are known to the VHF and with whom they have a direct contractual relationship. These users apply to use the archive and sign a non-disclosure document first to conduct research. If they seek to publish, they return to the Foundation for a license for permission to publish and/or to obtain broadcast quality video for use in an exhibit or documentary.
- ***Location-authenticated use,*** where the VHF knows who and where the users are. This use will be through the interfaces at identified museums, such as the Holocaust Memorial Museum in Washington, D.C. or the Museum of Tolerance Simon Wiesenthal Center in Los Angeles.
- ***Location-unauthenticated use,*** where the VHF knows where but not who the user is. This may be an exhibit in a museum or restricted use of the catalog and finding aids.
- ***Globally-unauthenticated use,*** where the VHF knows neither who nor where the user is; potentially the Internet at large. This level of use is not being permitted by the VHF yet, as the staff believe that much work will have to be done on the authenticated users first in order to understand how the materials might be used and how they should be accessed.

Other intellectual property concerns

The bulk of this section has dealt with copyright, although as we have just seen privacy concerns are important as well. Even when a work you plan to use is not protected by copyright law or does not have privacy issues associated with it, other intellectual property rights may come into play. The material may be protected by other laws, such as publicity, patent, trademark, and trade secret laws.

Databases

The availability of information is of paramount importance to maximizing the potential of the Internet. Legal protection of databases is therefore an important and timely issue. In some countries, databases are protected by copyright as original compilations of information, where their creation has involved effort. Until the ruling in the U.S. case of FEIST PUBLICATIONS, INC. v. RURAL TELEPHONE SERVICE CO., (499 U.S. 340)[4] in 1991 that only original creations were afforded copyright protection, some had assumed that creations which depended solely upon ‘the sweat of the brow’ were afforded protection. In this instance the Court held that as a telephone directory was simply an alphabetic listing it did not contain sufficient creative expression to be eligible for copyright protection. This judgment has been widely interpreted to imply that in the U.S. there is little copyright protection of databases of facts which exhibit no creative

originality. This is in line with most of continental Europe. In order for a database to warrant copyright protection, its author must have made some original effort in the collection, selection, and arrangement of material (e.g. by indexing terms or adding keywords). This protection is irrespective of whether the individual contents of the database are original and therefore copyrightable in themselves or include factual data (where the database protection will not prevent an individual from extracting them, short of copying the selection and arrangement of the database as a whole). Currently in the U.S., if the database is unprotected by copyright law, the contents of the database may be copied unless such acts are prohibited by contract or license. In Europe, however, the Database Directive adopted by the European Parliament in 1996 includes a right that prohibits (for fifteen years from the date of the database's creation) the extraction or reutilization of any database in which there has been a substantial investment in either obtaining, verifying, or presenting the data contents, without any requirement for creativity or originality. Databases are thereby afforded some element of sui generis legal protection. (See the Copyright office's 1997 Report on Legal Protection of Databases at http://www.copyright.gov/reports/ for a good discussion of contractual practices for accessing data.)

Watch this space: This is a hotly debated issue at the moment in the United States.

Registration

Registration provides a mechanism to strengthen the protection that copyright provides. Even though it is no longer necessary to use the copyright symbol © or print a copyright notice in order for your materials to be afforded copyright protection, doing so will ensure that the court will not accept a defendant's 'innocent infringement defense'. Similarly, while registering your materials with the Copyright Office is no longer a requirement, registration offers several benefits which are otherwise lost. These include the ability to bring an infringement case to court, to prove certain facts in a lawsuit, and to claim statutory damages and attorney's fees. Registration material is available in the U.S. at the Copyright Office, Library of Congress (http://www.copyright.gov/register/index.html).

The preparation of effective copyright notices depends upon knowledge of legal frameworks not only in the U.S., but in other countries. For example, according to the Buenos Aires Convention, in order for your material to gain copyright protection in Bolivia and Honduras, you must include the phrase "all rights reserved". While preparation of copyright notices is best done with appropriate legal advice, Benedict O'Mahoney's copyright website includes valuable guidance on how to draft such notices (http://www.copyrightwebsite.com/info/notice/notice.asp).

Technologies for copyright management and protection

Throughout this section we have discussed the need to strike a balance between protecting intellectual property and providing public access to it. In an effort to meet both these objectives there has been an increase in the design of electronic mechanisms which can mark or tag digital cultural assets and trace their usage. In addition to these technical mechanisms, the use of appropriate metadata can play an important role in making users aware of who owns the rights in material, the kinds of uses to which the users are allowed to put the materials, and the conditions that govern different uses (e.g. free to look at on your screen, but you must pay a fee if you wish to forward it to someone else). The encoded metadata might include a description of the object, details of the rights owner, as well as a unique registration number.

Digital Object Identifier (DOI)

The **Digital Object Identifier (DOI),** a system which has been proposed by the International DOI Foundation (http://www.doi.org) in conjunction with the Association of American Publishers and the Corporation for National Research Initiatives was launched in 1997 and is used by several publishers. It provides a unique identification number to digital objects (as one metadata element), allows rights holders to link to the users of the material, and enables automated copyright management for all types of media. By clicking on a DOI, the user is connected to a central directory where the Web address used by the publisher is located. This address is maintained by the publisher and may contain the content itself or further information on how to obtain the content and means of linking users to rights holders. As this might also contain a description of the object, it is important that DOI developments take into account the Dublin Core scheme and that the two communities work together towards harmonization of the two models (Bearman 1999).

Despite active research, currently available technologies for protecting assets have limited effectiveness and little general acceptance. They all aim to provide a mechanism recording and linking rights information to the digital object that is indelible and unmodifiable but perhaps also updateable and able to 'follow or accompany' the digital object when copied. Research for this Guide showed that most projects have chosen to not provide high quality archival or even medium quality resources online (e.g. Images of England, http://www.imagesofengland.org.uk), while they wait for digital rights protection technologies to become more effective, standardized, widely used, and affordable.

Link Box:

Some examples of advanced technologies for management and protection of digital assets include:

- **steganography**, which hides the information under a stegoimage, a cover image where the original is embedded (Johnson & Jajodia 1998, http://www.jjtc.com/pub/r2026a.htm; see also the infosyssec site on Cryptography, Encryption and Steganography, http://www.infosyssec.org/infosyssec/cry2.htm)
- **cryptography** which encrypts or scrambles the information so that it cannot be understood without a decoder, a pre-defined decryption key (for an introduction to cryptography, see the page by SSH Communications Security, http://www.ssh.fi/tech/crypto/intro.cfm)
- **digital wrappers**, protection systems which place digital content inside secure 'containers' which require a helper application and a key code to be opened. For an example see InterTrust's DigiBoxes, http://www.intertrust.com.
- **digital watermarking**, which is a digital signal or pattern inserted into a digital image and can be either visible or invisible. An early example includes the visible watermarks designed by IBM which were inserted in the images of the Vatican Digital Library Project (Mintzer et. al. 1996, http://www.research.ibm.com/journal/rd/mintz/mintzer.html). Although visible watermarks discourage theft and misuse, they can interfere with studying the original and decrease the overall quality of the digital object, so there seems to be a preference for invisible ones in the cultural and education sector. Even the most sophisticated digital watermarks however, can be overcome through overmarking, adding noise, removing the watermarking, and counterfeiting. For example, the Secure Digital Music Initiative (SDMI, http://www.sdmi.org) digital watermarks were defeated within three weeks. For more information on digital watermarking and steganography, see Fabien Petitcolas' page at the University of Cambridge Computer Laboratory, http://www.cl.cam.ac.uk/~fapp2/steganography/.

---

[1] Section 107 of U.S. Copyright law:
http://www.loc.gov/copyright/title17/92chap1.html#107

[2] Section 106a of U.S. Copyright law:
http://www.loc.gov/copyright/title17/92chap1.html#106a

[3] For further information on risk management see:
http://www.utsystem.edu/ogc/intellectualproperty/riskmgt.htm

[4] See http://www.law.cornell.edu/copyright/cases/499_US_340.htm for details of this case.

V. Digitization and Encoding of Text

Introduction

Digitized text is an important component of many cultural heritage projects, but the question of how to digitize it—what format, how much detail, what user activities to support—is a complex one. As with the digitization of other kinds of materials, you must consider a number of factors including the nature of the original material, the purpose of the digitized version, and the availability of relevant expertise, technical support, and funding.

It is also important to be aware of the purposes and limitations of the various digital formats available, to be sure not only that they suit your current goals and requirements, but also that they do not restrict your options in the future. Digital text formats vary considerably in the ease with which they can be converted to other formats, and in the variety of output methods they support. Proprietary systems such as word processing or page description formats (e.g. Microsoft Word, PDF) may be powerful and convenient tools for creating printed output, and may also allow for web publication of the results (via a "save as HTML" function), but if you need to move your data to another software platform you risk losing formatting and other information. Because such systems depend on the existence of proprietary software—whose licensing terms and very existence cannot be counted on over the long term—they are unsuited for archival purposes or for the creation of durable cultural resources.

In the past ten years, there has been a rapid growth in standards-based methods of text digitization using Standard Generalized Markup Language (SGML) and more recently its derivative, Extensible Markup Language (XML). These approaches avoid the problems of proprietary software, offering data longevity and the flexibility to move from platform to platform freely. There are now increasing numbers of tools for creating, editing, publishing, and manipulating textual materials encoded in this way, and this trend is likely to continue. Our treatment of digital text will therefore focus on standards-based methods of text digitization, which offers the best long-term solution to the needs of projects creating digital cultural heritage collections.

It may be useful to establish some basic terminology at the start. By "digitization" we mean any process by which information is captured in digital form, whether as an image, as textual data, as a sound file, or any other format. When speaking of the digitization of documents, the term may refer either to the capture of page images—merely a picture of the document—or to the capture of a full-text version, in which the document is stored as textual characters. In its most minimal "plain-text" form, a full-text version of a document may be simply that: the text of the document expressed as ASCII or Unicode characters and nothing more. Unlike a page image, such a document can be searched for particular words or phrases, but does not convey any information about the original

appearance or structure of the document. An "encoded" version of the same document will include additional information or "markup" of various kinds, expressing the document's structure, its formatting, or other information its creators wish to capture. Although strictly speaking, we can use the terms "markup" and "encoding" to refer to a wide range of added information—including word processor formatting codes or encryption—these words are now most frequently used to refer to SGML or XML markup. And although this kind of markup is usually applied to full-text documents, it is also possible to embed page images in an SGML or XML-encoded document structure, or to pair images with encoded information such as subject keywording, publication data, or administrative metadata.

Definition Box:

Definitions

***Page Image*:** A digital image of a page of text, captured by a scanner or digital camera, and expressed as a set of pixels in a format such as JPEG or TIFF.

***Encoding, markup*:** In this context, the process of adding information to a digital text by including markup (usually SGML or XML) which explicitly identifies structural and other features of the text. The term "markup" refers to the added information. In a broader sense, encoding may refer to any kind of added information or algorithmic transformation which, when applied to a data file, enables it to perform some special function.

***Keying*:** A process by which a person manually types, or 'keys,' text from source page images, original printed materials, photocopies, or microforms.

***OCR*:** Optical Character Recognition, a process by which software reads a page image and translates it into a text file by recognizing the shapes of the letters with various levels of accuracy that are difficult to predict. OCR-generated text tends to be described as either "uncorrected (or raw)" or as "corrected."

***Metadata*:** Strictly speaking, any data that is about other data; in this context, more specifically, the term usually refers to information describing a data file or document (for instance, publication information, revision history, data format, rights status).

***SGML*:** Standard Generalized Markup Language, an international standard (ISO 8879) since 1986, is a metalanguage which can be used to define markup languages.

***XML*:** Extensible Markup Language, a subset of SGML which was published as a W3C recommendation in 1998.

***TEI*:** The Text Encoding Initiative, an international consortium that publishes the TEI Guidelines for Electronic Text Encoding and Interchange, an SGML- and XML- compliant encoding language for the capture of literary and linguistic texts, widely used in the scholarly and cultural heritage communities.

***EAD*:** Encoded Archival Description, an SGML- and XML-compliant encoding language used for the capture of archival finding aids, widely used in the library and archival communities.

METS: Metadata Encoding and Transmission Standard, an XML-compliant standard for encoding a variety of metadata about digital library objects.

Having selected the materials you wish to digitize, (and having decided whether your text digitization strategy will include producing page images) there are several methods to choose from.

Page images are produced by scanning the text. Optical character recognition software reads such an image and creates a full-text version of the document by identifying individual character shapes and translating them into actual letters. Note that there are many factors affecting the accuracy of the OCR process, including the contrast of the original document, the fonts used, etc. Alternatively, if you have decided that page images are not needed, you can key text directly from a variety of sources. The formats available, and the handling policies pertaining to them, will dictate whether keying will be done from original materials, photocopies, or microforms. You can have the document keyed by project personnel or by a data capture service. Markup can be added during the keying process or subsequently. Increasingly, data capture services are willing to cater for specialized encoding needs, even using complex markup languages like those in the guidelines of the Text Encoding Initiative (TEI) and Encoded Archival Description (EAD); see the more detailed descriptions below and the URLs at the end of the section. It may be also possible to automate the encoding to some extent; how much will depend on several factors, including the consistency and predictability of your data, the level of encoding you require, and the amount of time and ingenuity you are willing to spend developing automated encoding tools. Some tools for this purpose exist already, and it is worth checking with existing encoding projects to see whether they may have something which will accomplish what you need. Finally, you may be able to obtain a digitized version of the materials you need from another project, such as the Oxford Text Archive or the University of Virginia Etext Center, either in plain text or with basic text encoding and metadata.

A few concrete scenarios may help to illustrate the differences between these various approaches, and the kinds of project goals each one is best suited to support. First, consider a fairly straightforward example: a project working with a collection of Victorian novels and poetry for which it owns copies of the materials to be digitized (the Victorian Women Writers Project, http://www.indiana.edu/~letrs/vwwp/, is an example of such a project). Since the materials are fragile and of historical interest, capturing page images makes sense both as a form of digital preservation, and in order to give readers access to an image of the original materials for research purposes. (However, given the uniformity of these texts and the general unremarkability of the pages, one can also imagine deciding against this step, particularly if the project did not own copies of the materials.) In addition, since these are literary documents, researchers will want to work closely with the text, so capturing a full-text version also makes sense. Adding TEI markup will enable readers to perform complex searches which take into account the genres and local textual structures which may be of interest. Since the page image has already been digitized, and since the materials in question are typographically regular, OCR is a good option for capturing the text, but since the project's audience is a scholarly one, careful proofreading will also be necessary to ensure that requirements for accuracy have been met. The question of how deeply to mark up the documents requires careful balancing of costs and desired function. Basic TEI markup can be applied nearly

automatically using scripts, but more detailed encoding requires additional staff, training time, and additional review for encoding consistency. The additional benefit to researchers would have to be substantial to justify this cost.

Consider another project in which a large quantity of heterogeneous materials must be digitized quickly, and made accessible for political and historical research (for instance, the Forced Migration Online collection at the Refugee Studies Project Centre at Oxford, http://www.rsc.ox.ac.uk/TextWeb/fmo.html). Retrieval is acknowledged as an important goal, but at the same time the nature of the individual documents does not seem to warrant complex markup; most users want to find documents relevant to their research, but are not interested in the texts' internal structure. Furthermore, given the volume of documents and the urgency of the project, costs and speed are important factors. The project's designers therefore decide to digitize all the documents as page images; since the documents vary so much and are likely to be unfamiliar, it is important for readers to see an image of the original. In addition, to support detailed retrieval at the topic level, they decide to OCR the text, but because of the volume and cost issues, they choose not to proofread each document. Since users are searching through such large volumes of material, the error rate of uncorrected OCR still provides acceptable accuracy when used only for indexing and query purposes. Users will see only the page images, not the full-text version, so the errors will not be visible. In addition, the project creates basic metadata for each document, to allow for accurate tracking and high-level retrieval.

These two kinds of projects have arrived at what look like fairly similar approaches (page images, text captured with OCR, and metadata), though for very different reasons. Consider now a third project, whose materials are medieval manuscripts to be digitized for literary and linguistic study. (See, for example, The Electronic Beowulf project, http://www.uky.edu/~kiernan/eBeowulf/guide.htm.) As before, there are likely to be important reasons to include images of manuscripts, but their uniqueness and fragility magnify the legal and logistical challenges. Because of the nature of the research to be supported, a full-text version is essential, but OCR is probably out of the question. Instead, the project needs to choose an appropriate method for keying in the text. For medieval manuscripts, this may require a scholar familiar with the manuscript to perform or oversee the transcription. Some materials may be readable enough by a non-expert to allow for keyboarding by a data capture service, or by locally trained encoders who are not subject specialists. Finally, to support literary and linguistic analysis, the project needs to use a detailed markup scheme such as the TEI, which will allow for the capture of features such as regularized or lemmatized readings of words, morphological and syntactic information, prosodic structures, textual variants, ambiguous readings, and so forth. With page images linked to this level of encoded full text, scholars can compare versions of the manuscript, check difficult readings against the image, search for particular linguistic features and compare their occurrence from poem to poem or from manuscript to manuscript. The collection is a full-fledged scholarly tool of extraordinary power, but it also requires considerable resources and expertise to create.

Projects planning to capture a full-text version of the document should also refer to the sections on OCR (Optical Character Recognition) and Keyboarding in Section VI, Images, in addition to the sections below.

Character encoding

Even a plain-text document without markup contains some very basic encoding of the character information in which the document is expressed. There are two principal character encoding systems which deserve brief discussion here: ASCII and Unicode.

ASCII—the American Standard Code for Information Interchange—was proposed by ANSI (the American National Standards Institute) in 1963, and finalized in 1968. It was designed to facilitate compatibility between various types of data processing equipment. ASCII assigns the 128 decimal numbers from 0 to 127 to letters, numbers, punctuation marks and common special characters. This is commonly referred to as the 'low ASCII character set'. Various extensions to the ASCII character set have been created over the years to assign the decimal numbers between 128 and 255 to special, mathematical, graphical and non-Roman characters. If you wish to provide texts in a 'plain text format', i.e. with the file extension "txt", then you must use the low ASCII character set. The extended ASCII character set has limitations that do not apply to the lower set and must be used with more caution. There is also more than one extended character set (IBM and Microsoft each have their own) and this diminishes interoperability.

There is some argument to be made for always providing a plain text format without markup, whatever other encoding scheme is used, especially in the creation of electronic corpora. There is not much extra work involved in making the texts available in this simplified form, especially as plain ASCII texts are often the most common starting point for creating other encoded text. Certainly encoded texts (for instance, SGML/XML, COCOA, or other formats) should be created and stored in plain ASCII rather than in a proprietary system such as a word-processing format, to guarantee their longevity and platform-independence. Projects such as the Thesaurus Musicarum Latinarum (TML), where Latin musical notation has been encoded using the low ASCII character set, demonstrate that encoding complex features, while maintaining a high degree of interoperability and longevity, can be achieved with ASCII characters. This high level of interoperability is particularly important for the TML because its small but highly dispersed user base uses a wide variety of hardware and software. Furthermore, the small file size of ASCII text files means the output of the project can be distributed quickly and easily.

Unicode is an international standard (implementing ISO 10646) which was developed to overcome problems inherent in earlier character encoding schemes, where the same number can be used for two different characters or different numbers used for the same character. Unicode assigns a unique number to each character; in addition, it is multi-platform, multi-language and can encode mathematical and technical symbols. It is especially useful for multilingual texts and scripts which read right to left. Originally

there were 65,000 characters available; now there are three encoding forms which can be used to represent over 1,000,000 characters.

Link Box for character encoding:

Unicode: http://www.unicode.org/

ASCII: http://www.asciitable.com/

Other formats which were used in the past are worth remembering:

- EBCDIC = Extended Binary Coded Decimal Information Code, the 8-bit character coding scheme used by IBM mainframes and minicomputers.
- Multicode: (see Mudawwar 1997)

Text markup

Text markup is a fundamentally interpretive activity, in ways that are both powerful and challenging. With a well-designed encoding language, you can express the facts and insights which are most useful for your purposes in working with the text. However, the interpretive quality of text encoding should not be taken to mean that encoded texts are purely, or merely, subjective creations. The most widely-used humanities text encoding languages such as TEI or EAD have been developed by particular communities whose needs and methodological assumptions they reflect.[1] These languages are designed to allow the representation of the significant information these communities want to capture. This information includes not only basic facts—or assumptions so deeply shared that there is no disagreement about them—but also significant interpretive statements, such as those of critical editing or archival description. One important role of the encoding language is to enable such statements to be made in a rigorous and consistent way, according to the practices of the community within which they are to be used. Text encoding thus makes it possible to bridge the gap between local research and insight and the discourse of the larger community, and to articulate interpretative statements in a way that is broadly intelligible.

The example given here shows a historical letter encoded using the guidelines of the Model Editions Partnership (MEP, http://mep.cla.sc.edu). As with all SGML-encoded documents, each textual element is enclosed within tags (set off by angle brackets) which mark its beginning and ending. The encoded text begins with a header (here abbreviated for simplicity) which contains the document metadata. Following this, the body of the document (encoded as <docBody>) contains first a heading, dateline, and salutation, and then the main text of the letter encoded as a series of paragraphs. At the end is a postscript with its own dateline. The elements and their grouping reflects the interests and research needs of documentary historians, for whom the MEP encoding scheme is designed.

Example Box:

A Sample Fragment of an SGML-Encoded Text (some encoding omitted for clarity)

```
<doc>
<mepHeader> ... </mepHeader>
<docBody>
<head>To <addressee>Martha Laurens</addressee></head>
<dateline>
<place>Charles Town</place>, <date>August 17, 1776</date>
</dateline>
<salute>My Dear Daughter</salute>
<p>It is now upwards of twelve Months ...</p>
...
<p>You will take care of my Polly too ...</p>
<signed>your affectionate Father</signed>
<ps>
<dateline><date>19th</date></dateline>
<p>Casting my Eye over ...</p>
</ps>
</docBody>
</doc>
```

The balance between disciplinary constraint and local expression is managed in SGML and XML through the document type definition (DTD), which is a formal statement of the tags permitted in the encoding system (i.e. what textual structures it is capable of naming) and how they may be nested. Different DTDs handle this balance very differently, depending on their intended purposes. Structurally, DTDs may be quite strict, specifying with great precision the order and nesting of tags, or they may be quite lenient, allowing greater latitude to the encoder. Similarly, in their definition of a tag vocabulary, DTDs may provide for very nuanced distinctions between features, or they may use fewer, more generally defined elements. Finally, DTDs may be constructed in an attempt to anticipate and codify all the possible encoding situations in the domain they cover, or they may instead provide methods for the encoder to handle unforeseen circumstances. All of these approaches have their potential uses, and in choosing or designing an encoding system it is essential to understand the nature of your own material and goals,

so that you can choose appropriately. Although a structurally strict encoding system may seem to limit one's options unnecessarily, in fact such a system can be valuable in constraining data entry and in ensuring that identically structured documents are all encoded alike. A more lenient DTD is superficially easier to work with, but through its flexibility it also opens up the likelihood of inconsistency and of time-consuming debate about which option to choose in a given circumstance. A DTD with a rich lexicon of tags may be essential for describing certain kinds of textual features in detail, but is an encumbrance when only a simple encoding is required.

DTDs and the encoding systems they represent can often be adapted by the individual project to suit local needs, and this is particularly true (and may be particularly useful) in the case of large, multi-genre DTDs like the TEI. However, one important function of the DTD is to allow for comparison and interoperation between collections of encoded data. By changing it, you in effect secede from the encoding community of which that DTD is the expression, and you diminish the possibility of using common tools for analysis, display, and retrieval. Some encoding systems are designed to be extensible; for instance, the TEI provides an explicit mechanism by which individual projects may define TEI-conformant tag sets which are adaptations of the TEI encoding scheme. When done with care—and preferably in concert with other projects with similar needs—such adaptations may improve an existing encoding system while avoiding the disadvantages described above. For projects dealing with highly idiosyncratic data, or projects attempting to capture features for which no encoding system exists, adaptation may be simply unavoidable. In such cases, you should be prepared to think through your changes carefully and document them thoroughly.

SGML and XML

Although there are a number of encoding systems which have been developed for humanities computing use over the past few decades—many of them still in use—Standard Generalized Markup Language (SGML) and its derivative, Extensible Markup Language (XML) deserve particular attention here both because they are so widely used and because they should be. As international standards,[2] they receive a level of attention from software developers and from the standards community which guarantees their comparative longevity, and because they are non-proprietary, they can be used to create archival collections which are free from software or hardware dependencies, and hence less prone to obsolescence.

Strictly speaking, SGML and XML are metalanguages: systems for defining encoding languages. Because they provide a standardized method for specifying things like how a tag is delimited or how its structural definition is written, software written to this standard can be used with documents encoded with any SGML or XML-conformant encoding language, regardless of the particular tag set or the kinds of data they contain. The most significant text encoding systems for cultural resources—TEI, EAD, CIMI, METS, and others—are all written in SGML and XML, as is the ubiquitous HTML. There also exist SGML and XML versions of data standards like MARC.

The advantages of SGML and XML, as suggested above, stem partly from their status as international standards. In addition, because this kind of encoding allows for the complete separation of structure and presentation, SGML/XML-encoded documents can be repurposed or used as a base format from which to derive specific versions for different purposes: word-processing files for printing, HTML for delivery on the web, Braille output for the visually disabled, and so forth. SGML/XML encoding is particularly valuable for the kinds of cultural heritage work covered in this report, because it permits the description of the text's constituent parts in terms which are meaningful for retrieval and intellectual analysis. We might express this whimsically by saying that an encoded document “knows” what its own parts are in the same way that a scholar or reader does: concepts like “heading” and “quotation” and “poem” and “author” are accessible as primary terms of analysis (assuming they are part of the encoding language used). Furthermore, unlike some of the earlier encoding languages that were designed to avoid verbosity at all costs, SGML/XML encoding is actually fairly easy to understand once the eye becomes accustomed to seeing the tags. Encoding languages like TEI and EAD use tag names which are expressive of their function—<note>, <author>, <name>, <list>, <quote>, and the like—and because they represent the ideas people actually have about documents, they quickly become intelligible even to the untrained reader.

The disadvantages of SGML and XML have largely to do with their formalism as data structures. Because of their requirement that all documents be expressed in the form of a hierarchy or tree of nested elements, they are not ideal for representing truly non-hierarchical materials (for instance, sketchbooks). Although XML cannot simultaneously represent multiple hierarchies in the same document (and SGML can do so only with great difficulty), these occur so frequently that systems have been developed to handle most ordinary cases, and in practice this is not usually an obstacle to the use of SGML or XML, only a design consideration.

While SGML in its many applications—HTML, TEI, EAD, and others—is widely used by cultural heritage projects, we are entering a period of transition where XML is becoming more widespread. XML is in effect a streamlined version of SGML, in which some features of SGML which make it unnecessarily complex to implement have been eliminated.[3] Like SGML, XML is a metalanguage which can be used to define specific encoding languages, and all of the encoding languages discussed here are now available in an XML version. However, rather than abandoning SGML in favor of XML, projects seem to be using both. With the growing availability of XML software for publication and browsing of online documents, XML has become a central component in the delivery of cultural heritage materials, and will only become more so in the future. Several factors are particularly significant in this shift:

- because XML tools are easier to write, there is already significantly more software for processing, publishing, and managing XML data, and these tools are not confined to the business and industrial markets. With the rise of XML, cultural heritage projects will thus find an increase in affordable software that meets their digital publication needs. One extensive though technical resource for information on these kinds of tools is the Cover Pages, maintained by Robin

Cover (http://www.oasis-open.org/cover/sgml-xml.html). The TEI also maintains a list of SGML and XML tools at http://www.tei-c.org/Software/.

- publication of XML is already simplified by the availability of tools to transform XML-encoded documents from one encoding scheme to another (e.g. from TEI to HTML), but in addition web browsers are now being developed which can read XML documents and style sheets, and hence can display documents from encoding systems other than HTML.

- XML makes it easier for users to use local style sheets (XSL and CSS) to display information the way they want it. It also makes it possible for document publishers to derive multiple encoded versions of a given document from a single master XML file, using XSLT stylesheets. This feature makes it much easier than before to create and manage versions adapted for different kinds of user needs (for instance, the visually impaired) or different delivery strategies.

Unlike SGML, XML documents do not require a DTD. For publication purposes, the lack of a DTD is not much of a concern as long as the document is well-formed (that is, as long as its elements all nest within one another and as long as they all have start-tags and end-tags). An XML style sheet (XSL) can still format the document and browsers can still display it without checking against a DTD to see that the document is valid. However, for production purposes, working without a DTD is not advisable, since it makes it impossible to check documents for consistency of encoding. An alternative to DTDs is XML Schemas, which are another way to specify and validate the structure of your documents. XML Schemas are currently under review by the W3C, and offer some advantages over DTDs, such as typing of elements, grouping of declarations, and inheritance of properties. For more details, see http://www.w3.org.

Developing a Document Type Definition

One of the first steps for new text encoding projects is to identify the encoding system and specific DTD to be used. As suggested above, this process also involves articulating the project's methodological commitments and audience, as well as its more general area of focus. There are a number of encoding languages that fall within the domain of the cultural heritage community, but their range of usefulness does not overlap by much. The TEI DTD is primarily intended for full-text encoding of humanities and scholarly texts; the EAD DTD addresses the needs of the archival community for encoding finding aids and guides to collections. At this point, projects that deal primarily with artifacts, works of art or audio-visual material are far less well served.

It is probably clear by now that if you can use an existing DTD for your materials, you probably should. However, if your material falls outside the realm of existing encoding systems, you may need to develop one yourself. Projects such as the Oriental Institute of Chicago, whose text material does not fall within the broad western text canon around which TEI and the majority of DTDs have been designed, must either develop their own

or await developments from others. While simple DTDs are easy to create, developing an encoding system that will fully represent your materials may require considerable research and time. The complexity of the task increases with the heterogeneity of the collection and the level of detail you wish to represent. For a project with a large number of texts with variable structures and features, it can take many years of development, application and refinement to produce a DTD that meets all of its requirements. This is by no means an impossible task, and important projects like the William Blake Archive (http://www.blakearchive.org/public/about/tech/index.html), The Orlando Project (http://www.ualberta.ca/ORLANDO/), the Electronic Text Corpus of Sumerian Literature (http://www-etcsl.orient.ox.ac.uk/project/sgml-xml.htm), and others have taken this approach. However, the implications for a project's funding, staffing, and training as well as the time-scale for deliverables must be taken into account.

Definition Box:

Definitions

DTD: Document Type Definition, the formal set of rules that define the elements that may occur within an encoded document and their structural relationships (their relative order and nesting)

Content model: A component of a DTD, giving the structural definition of a particular element

Occurrence indicators: Within a content model, the occurrence indicators show how often a given element may appear (once only, at least once, or any number of times) and whether it is required or optional.

#PCDATA: Parsed Character Data, i.e. words and spaces but not tags.

Tag: an individual piece of encoding that marks the start or end of a textual feature, set off from the document's content by special characters (in practice, usually angle brackets: <tag>).

Element: A textual feature within an encoded document, including the start-tag, the end-tag, and the encoded content: <name>John Smith</name>

Attribute: a modifier to an element, almost as an adjective or an adverb modifies a noun. Attributes come in many varieties and may be used to indicate the type of element (for instance, <name type="person">), its location in a series, a link to a related element, the language of the element's content, an alternate reading, and a wide variety of other kinds of information.

Entity reference: a special character sequence which is used as a placeholder for some other text. Entity references are often used to encode characters which cannot be typed directly in ASCII, such as accented characters, ornaments, or non-roman alphabets. They may also be used as a placeholder for boilerplate text. In SGML and XML, entity references typically begin with an ampersand and end with a semicolon, e.g. é Entity references can also be used to point to external files, such as page images, that can be referenced in the markup and displayed as if they were embedded in the text.

The example below shows a simple DTD fragment that describes a very basic encoding for poetry. Each permitted element is declared together with a specification of what it may contain, and in what order. The occurrence indicators (question mark and plus sign)

indicate whether the element in question is optional, and how many times it may occur. The commas separating the elements indicate that the elements must occur in this order. Thus the first line of this DTD specifies that a poem may start with an optional heading, followed by at least one or more <lg> elements, and ending with an optional closer. The second line indicates that the <poem> element also has a required type attribute, which provides a way of identifying the kind of poem more specifically. In this case, the DTD defines a list (unrealistically brief) of possible values, although it is also possible to leave the values unspecified.

Most of the other elements in this DTD are defined so as to contain simply #PCDATA, or Parsed Character Data (in other words, any valid character). However, the <lg> (line group) element has a slightly more complex content model. It may contain <l> or <lg> elements; the vertical bar indicates that either one may occur. The plus sign means that one or more of the group of elements it modifies (in this case, <l> and <lg>) must occur. The net result, therefore, is that an <lg> element may contain one or more verse lines, or one or more nested <lg> elements, or a mixture of the two. It may not contain naked characters (without an enclosing element), nor may it be empty.

Example Box:

A Simple XML DTD Fragment

```
<!ELEMENT poem (head?, lg+, closer?) >
<ATTLIST poem type (sonnet | stanzaic | irregular ) #REQUIRED >
<!ELEMENT head (#PCDATA) >
<!ELEMENT lg (l | lg)+ >
<!ELEMENT l (#PCDATA) >
<!ELEMENT closer (#PCDATA) >
```

The encoded example text that follows represents one of many possible valid documents conforming to this DTD. Equally valid would be a poem consisting of a single line group containing a single line, without a heading or a closer. When designing a DTD, it's equally important to consider the elements you wish to be able to omit, and the elements you wish to require. In other words, you need to decide not only what constitutes the minimum valid document, but also what constitutes the greatest variation you may need to accommodate.

Example Box:

A Simple XML Encoded Document:

```
<?xml version="1.0" encoding="US2" standalone="yes" ?>
<poem type="stanzaic">
<head>The Clod and the Pebble</head>
<lg>
<l>Love seeketh not itself to please, </l>
<l>Nor for itself hath any care, </l>
<l>But for another gives it ease, </l>
<l>And builds a heaven in hell's despair. </l>
</lg>
<lg>
<l>So sung a little clod of clay, </l>
<l>Trodden with the cattle's feet, </l>
<l>But a pebble of the brook</l>
<l>Warbled out these metres meet: </l>
</lg>
<l>Love seeketh only Self to please, </l>
<l>To bind another to its delight, </l>
<l>Joys in another's loss of ease, </l>
<l>And builds a hell in heaven's despite. </l>
</lg>
<closer>William Blake, Songs of Experience</closer>
</poem>
```

The remainder of this section will discuss in more detail some text encoding languages of particular relevance for cultural heritage materials.

HTML

A brief discussion of HTML is warranted here if only because it is so widely used and so familiar. HTML is essentially a formatting and display language for the web, designed as a small set of tags for simple hyperlinked documents, and as such its value as a form of descriptive markup is extremely limited. It lacks the vocabulary necessary to describe many of the basic features of cultural heritage materials—most significantly their metadata, their genres, and their textual structure. In cases where HTML is adequate for describing such materials (because the materials themselves or the representations desired are extremely simple) a simple TEI-based DTD would be nearly as easy to use and much more upwardly mobile. XHTML offers some improvements over HTML. Although it offers no greater descriptive power (since it provides the same tag set as HTML 4.0), it does allow for the enforcement of XML compliance such as being well-formed, allowing validation against DTDs or schemas, and extensibility through the formal definition of new modules. For projects that choose to use some form of HTML, XHTML will at least offer better interoperation and increased delivery options (for instance, to the increasing variety of web-enabled devices such as mobile phones and hand-held computers).

Projects that create SGML-encoded texts still rely heavily on HTML because SGML encoded texts cannot be viewed online by most web browsers, and there is a grave shortage of SGML-aware software at this time. Projects such as the William Blake Archive or the Victorian Women Writers Project have developed tools to convert SGML documents into HTML for viewing on the web. Others, such as the Women Writers Project, use commercial software that performs the translation to HTML dynamically. With the advent of XML, web publication is likely to become much more straightforward, and conversion to HTML as an intermediate will become unnecessary.

Definition Box

HTML

- HTML is an SGML application; HTML 4.01 is the latest standard
- HTML is designed to instruct a browser how a page should be rendered, although the lack of standards for browsers has led to considerable variability in the actual on-screen appearance
- HTML is poor at representing information about a text's structure
- Certain HTML features are browser-dependent
- HTML can be generated from SGML/XML for display

XHTML

- XHTML is a reformulation of HTML as an XML 1.0 application; XHTML 1.1 is the latest W3C recommendation
- XHTML can operate in HTML 4 conformant browsers
- Because XHTML documents are XML conforming, they are readily viewed, edited, and validated with standard XML tools

TEI (Text Encoding Initiative)

For projects creating full-text resources, the TEI Guidelines[4] are the predominant choice. The Guidelines have been adopted by a large number of projects representing a range of different kinds of text, and have proved highly adaptable to local requirements. Among the projects surveyed, the use of TEI DTDs in encoding texts is one of the clearest cases of the adoption of standards for a particular type of material. Their use indicates the close match between the TEI's goals in creating the guidelines and the goals that text encoding projects had in mind when creating their texts:

- to represent textual features for research
- to have simple, clear and concrete encoding standards
- to provide a common core of shared textual features
- to enable user-defined extensions or constraints
- to conform to existing and emergent standards
- to allow multiple parallel encoding of the same feature
- to provide rigorous definition and efficient processing
- to allow the encoder to define the level of markup detail
- to provide for documentation of the text and its encoding

At the same time, most projects have found either that the full implementation of TEI is unnecessary, or that the benefit did not justify the extra time and intellectual effort required. Many have turned to the TEI Lite DTD, a simplified view of the full TEI DTD. [5] The purpose of TEI Lite—meeting 90% of the needs of 90% of users—seems to be borne out in practice, and TEI Lite has become the common starting point for a large number of text encoding centers and projects, including the Virginia Etext Center and the Michigan Humanities Text Initiative. While an understanding of the full TEI Guidelines is still desirable, not least for deciding what elements can be ignored, the use of TEI Lite is recommended as a starting point for good practice in text encoding. It is always possible to add further layers of detail at a later stage, if your needs change.

The basic structure of a TEI-encoded document is very simple. Every TEI document must begin with a <teiHeader> element, which contains the document metadata. The header may be very simple, but can also accommodate detailed information about the electronic text's publication, source, subject matter, linguistic characteristics, and revision history. Following the header is a <text> element which in turn contains <front>, <body>, and <back>, which in turn contain elements. In addition to accommodating all of the typical features of texts—paragraphs, lists, headings, the various components of poetry and drama, names, dates, quotations, bibliographic citations, and so forth—the TEI Guidelines also provide for more specialized encoding of features such as prosodic

structures, morphological analysis, subject keywording, and similar features which are useful for various kinds of scholarly textual research.

EAD (Encoded Archival Description)

Although the main thrust of DTD development has been in the direction of humanities and scholarly texts, several other DTDs have been developed to cater for heritage institutions with different text encoding requirements. The most significant of these for the archival community has been the Encoded Archival Description (EAD) DTD.

The EAD DTD began as a cooperative venture in 1993 at the University of California, Berkeley. It aimed to develop a non-proprietary encoding standard for archival finding aids that would include information beyond what is provided by traditional machine-readable finding aids, such as MARC.

The project chose SGML as the most appropriate encoding language, as its document type definition (DTD) concept makes it ideal for the consistent encoding of similarly-structured documents, the key to successful electronic processing. An analysis of the structural similarities of finding aids helped construct the initial FINDAID DTD. This simplified, improved and expanded access to archival collections by linking catalog records to finding aids, enabling the searching of multiple networked finding aids and keyword access. The release of version 1.0 of the EAD DTD was delayed until 1998 in order to make it compatible with the emerging XML.

EAD documents consist of two major parts. The first part is the <eadheader> element, which contains metadata about the finding aid and its encoded representation. The second part is the <archdesc> element, which contains the information about the archival materials described.

The EAD header was modeled on that of the Text Encoding Initiative (TEI). It consists of four elements (some of which are further sub-divided):

- <eadid> (i.e. EAD identifier): provides a unique identification number or code for the finding aid
- <filedesc> (i.e. file description): contains much of the bibliographic information about the finding aid
- <profiledesc> (i.e. profile description): records the language of the finding aid, who created the encoded version and when
- <revisiondesc> (i.e. revision description): summarizes any revisions made to the EAD document or its encoding.

The uniformly ordered elements in the <eadheader> make searches more predictable. Such searches can filter large numbers of machine-readable finding aids by specific

categories such as title, date and repository. The <eadheader> is obligatory, so archivists are forced to include essential information about their finding aids that were not recorded in paper form. The optional <frontmatter> element can be used to create title pages that follow local preferences.

Because finding aids generally describe material at several different, but related levels of detail, these unfolding, hierarchical levels are represented within the <archdesc> element. The <archdesc> provides for a descriptive overview of the whole unit followed by more detailed views of the parts. The data elements that describe the whole unit are gathered together under a parent element called <did> (descriptive identification). These <did> elements are the key to good description as they facilitate retrieval of a cohesive body for discovery. Once the high (or unit) level of description is complete, the component parts can be described using the Description of Subordinate Components or <dsc> tag, at whatever level of detail is appropriate for the collection and the resources available.

Dublin Core

While not an encoding system in its own right, the Dublin Core deserves a reference here as part of good practice in creating encoded metadata. The Dublin Core Metadata Element Set defines a set of 15 essential metadata components (for instance, author, title, format) which are broadly useful across disciplines and projects for resource discovery and retrieval. These components can be used to add metadata to HTML files (using the <meta> tag) but can also be used in other contexts to create basic metadata for a wide range of digital resources. Dublin Core does not provide for detailed administrative or technical metadata, and as such is largely suited for exposing resources for search and retrieval, rather than for internal resource management and tracking. In addition, since its goal is to be simple and broadly applicable to a wide variety of resources, it does not provide for the kind of highly structured metadata about specific document types that TEI and EAD offer. Although projects using these encoding systems will probably not need to use the Dublin Core, they may find it useful to be aware of it as a possible output format for distributing metadata about their resources.

METS

The Metadata Encoding and Transmission Standard (METS) is an XML-based encoding standard for digital library metadata. It is both powerful and inclusive, and makes provision for encoding structural, descriptive, and administrative metadata. It is designed not to supersede existing metadata systems such as Dublin Core or the TEI Header, but rather to provide a way of referencing them and including them in the METS document. As a result, it is an extremely versatile way of bringing together a wide range of metadata about a given digital object. Through its structural metadata section, it allows you to express the relationships between multiple representations of the digital object (for instance, encoded TEI files, scanned page images, and audio recordings), as well as relationships between multiple parts of a single digital representation (for instance, the

sections of an encoded book). Its administrative metadata section supports the encoding of the kinds of information projects require to manage and track digital objects and their delivery: technical information such as file format and creation; rights metadata such as copyright and licensing information; information about the analog source; and information on the provenance and revision history of the digital objects, including any data migration or transformations which have been performed. METS is a very recently developed standard but is well worth watching and using.

Link Box:

Links to Useful Resources for digital representation and markup of text:

- TEI: http://www.tei-c.org
- EAD: http://lcweb.loc.gov/ead
- Dublin Core: http://dublincore.org
- METS: http://www.loc.gov/standards/mets/

---

[1] Arguably, in other domains such as industrial or computer documentation, where text encoding languages are intended to govern the creation of new digital documents rather than the representation of existing ones, encoding cannot by its nature be interpretive, since the author and encoder operate with the same agency (even if they are not the same person).

[2] Strictly speaking, while SGML is an international standard (ISO 8879), XML is only a recommendation from an international consortium (the World Wide Web Consortium). In practice, this is a distinction that makes little difference.

[3] It is important to note that the greatest impact of these differences is on software design; to the encoder and the end user, the change from XML and SGML is not difficult to make.

[4] The TEI Guidelines for Electronic Text Encoding and Interchange (P3, the third release but the first official publication) were published in 1994. The latest version, P4, adds XML compatibility and was published in March 2002.

[5] The TEI DTD is designed as a set of modules which can be combined in various ways to accommodated many different types of texts. Thus there is no single TEI DTD, but rather a set of DTDs that represent the various combinations of the modules. The TEI Lite DTD is a simple combination of the most widely required TEI modules.

VI. Capture and Management of Images

Introduction

Images have tremendous power on many levels and in many contexts. Before the advent of the Web, the online world was largely limited to textual data, but with the growth of the Internet, the demand for images is huge and continues to grow. In addition to the compelling visual pleasure they offer—which puts them in high demand in the culture generally—they also carry dense cultural meaning and are being used increasingly for research, teaching, and learning.

As a creator and distributor of digital images you have many factors to think through.

As with other kinds of materials, these include the selection of objects and visual documentation for digitization, the intended purpose of the digitized items or collection(s) and their expected users and uses, as well as the availability of relevant resources. Your approach will be determined by a number of considerations: the nature of your material, the needs of your users, and your available financial resources and human expertise. Overall, you should remember that your objective is to create a high quality digital resource that will remain useful in the long term. This resource should have as its cornerstone an archived digital master that can be considered "use-neutral" and can serve as the parent to many children that are used in print, web, video and other forms of reproduction.

Other sections in this *Guide* provide information that bears directly on the issues in this section. Section II provides a broad contextual view on overall project planning, including many of the larger decisions that affect image capture. Section III gives guidance on selecting materials, which will play a crucial role in determining what kinds of image capture techniques are appropriate. Section VIII offers detailed information on quality control and assessment, and Section XII discusses user needs evaluation, which can help you decide issues such as the required image quality. With this information as a backdrop, an overview of the steps to be taken will include the following:

- Investigate the formats of the source material (works on paper, original paintings, slides, negatives, transparencies, photoprints, maps, postcards, etc., as this will affect your choices of method and equipment and your need for specialized expertise. It is likely that the project contains multiple types of material that need to be evaluated both separately and as a group when planning the scope of the project, the equipment needed, and the most effective work flow.

- Establish an understanding of the characteristics and properties of your source material (its fragility, size, mobility, access, etc.) as this will also affect method, equipment, costs and expertise.

- Assess the way it conveys information (how important is color, line, detail, text?) and the kinds of metadata that can or must accompany it.
- Decide whether you will digitize in black-and-white or color (a variety of bit-depth decisions need to be made).
- Decide on whether you have the expertise needed in-house, whether you need some training or whether you will outsource the job.
- Decide on file format, size, storage and management.
- Define benchmarks to ensure image capture consistency across the collection.
- Test a sample of images for capture, display, and print output.
- Decide on standards for documentation (metadata) of the digital files.
- Clarify decisions about sustainability of the project and preservation of the digital files.

Finally, we emphasize that this section should be read alongside the section on Quality Control and Assurance, which is an integral element of the creation process.

Planning for Image Capture

Choice of Source: Original versus Intermediate

During the selection stage, you will have chosen the items to be digitized on the basis of their content or information value as well as for their potential importance to a set of users. At this stage in the planning you should also consider whether to digitize from the original object or from an existing surrogate, for instance if the original is inaccessible or could suffer damage from handling during the process. Examples of derivative formats are photographic prints of original art works, microfilm of newspapers, and slide images of objects.

Frequently, the physical size and condition of the object will determine this question. If an original artwork is oversized, then a flatbed scanner will not work. In addition, some objects should not be placed in contact with glass. For example, fragile work, determined as such by a conservator or preservation specialist, should not be put in contact with a scanner platen or a glass plate holding materials flat on a camera copystand. It may be possible to digitize such objects if a high-end digital camera is in place, with the requisite equipment, such as book cradle and cool lights. Even “cool lights” may emit harmful ultra-violent radiation or destabilize the temperature and humidity in the room, and if this will put the objects at risk you may need to look into the latest generation of electronic flash-based, area array digital cameras. If such equipment is not available, digitizing from

a film intermediary may be a possible alternative. With any intermediary there will always be some information loss, so digitizing from the original is advisable when the highest level of quality, per item, is needed, in cases such as preservation replacement. It should be noted that digitizing from the original can significantly raise costs—not only for capture, but also for preservation—which is why intermediaries are used so widely in large-scale, production-level digitization projects.

Intermediaries pose their own problems, including poor production and scratched, faded, or out-of-focus images, all of which affect the quality of the digital product and need to be closely monitored through the quality control and assurance process. However, staff at some projects included in our survey, for example those at the University of Michigan, photograph all the originals and digitize the resulting 4" x 5" transparencies. This approach is not common among the major projects, and most will digitize from the originals. It should be noted, though, that it is better to do a good job of digitizing a well-made intermediary than a merely adequate job of digitizing original images.

Before deciding to use intermediaries, it is worth considering the following factors:

- Condition of the object: is it rare and fragile? will it be damaged by exposure to intense or unfiltered light?

- Physical size of the object: is it too large to be managed on the available equipment?

If the answer to these questions is yes, then you should consider using a film intermediary.

Quality and Resolution

Decisions as to what resolution, bit-depth or method of tone reproduction, and image format to use for capture, storage and delivery can only be made after project staff have completed an assessment of the material, the needs of potential users, the available technology, costs per image and likely financial support.

The key questions will be:

- What information content do you wish to convey and in what format(s) – screen image, print, or both?

- What functionality does the end product need to support? For instance, does it need to accommodate art history research, which might require scrutiny of minute details such as brush strokes at high magnification? Or is it intended for less exacting audiences where download speed is more important?

- What capture standards and equipment will achieve this?

While you could produce a matrix and try to balance all these parameters, experience shows that you will inevitably need to make trade-offs. Although local circumstances such as resource constraints, long-term plans for the material, and other factors will dictate what trade-offs are appropriate, other things being equal it is probably a good rule of thumb to reduce quantity and maintain quality. A higher-quality digital object offers greater flexibility for future use, since you can always derive lower-quality versions from it, whereas if your needs become more demanding, a lower-quality image will be inadequate. This consideration may carry additional weight if you are dealing with rare or perishable materials, where you may not have a second opportunity to digitize in the future.

There is no set resolution for any individual resource or collection. Each project must identify the minimum level of quality and information density it requires for its digital surrogates. A detailed explanation of sampling rates and other measures of information density can be found in the appendix on digital data capture. Briefly, resolution is usually expressed in dots or pixels per inch (dpi or ppi) and measures the density of sample—the number of information samples taken per unit of area— that is captured by the scanning equipment. Generally speaking, the higher the ppi, the more detail being captured. The sampling depth, or bit-depth, measures how much information is captured in each sample: for instance, how many colors or levels of grayscale. The higher the sampling depth, the more subtle the gradations across the image, and the larger the resulting file size. The sites interviewed for this *Guide* used a range of sample rates from 300 dots per inch (dpi) to 600 dpi for their master image files. Lower resolution images were often derived from the master file to produce distribution copies with smaller file sizes for ease of download. The most widely adopted format for storing preservation quality digital masters is uncompressed TIFF. Derivatives, in such file formats as JPEG, are used for access and delivery versions of the image. The resolution of these derivatives generally ranges between 72 ppi and 150 ppi. The resolution appropriate for a high-quality image of the source material should be determined by the size of the smallest significant detail that will need to be visible, given the intended use of the image.

Images can be either reflective (e.g. photographs or drawings) or transmissive (e.g. transparencies or negatives). With reflective material, light is bounced off the surface of the image and its intensity measured using light sensitive diodes. In the case of transmissive material, light is instead passed through the original source. There are three constraints on the highest resolution you can use in scanning both reflective and transmissive material: the maximum sampling capabilities of the scanning device; the size of the files you can manage; and the level of information you actually need. An example of this last point involves old postcards, which were often printed on poor quality paper using cheap processes. If scanned at high resolutions you capture the texture of the paper, which can obscure other information in the image.

Remember that before digitizing you still need to answer questions about the format of the material, its size (or size range if in a batch); whether you need color or black and white (see the discussion of this question in the cost-benefit analysis section of Section

III: Selecting Materials); what is the smallest significant detail in the image that you wish to capture (fine line drawing, detailed photographs, fine art); and to what range of uses will the digital objects be put?

Benchmarking

The greater the detail that needs to be captured, the higher the resolution required. Many projects use benchmarking to identify the best resolution for capturing a given image. Benchmarking will identify the smallest essential detail by investigating the attributes of the source and will address the issues of whether the objective is to convey the informational content of the original or to convey information about its creation, such as brush strokes or engraving methods.

One way to establish a benchmark is to take a sample of the material, perhaps that with the greatest detail, and scan at a variety of resolutions. Show these scans, using the medium on which they will be chiefly viewed in practice, to a sample of users and staff to identify which sampling rate best meets the needs for the digital object. This process will be most effective if you take into account the kinds of equipment your users will actually be using: will they have access to high-end monitors and fast broadband connectivity? Once you have identified the optimum resolution then you can set the specifications for that collection. An alternative to 'evaluative benchmarking' is to use a Quality Index (QI). The Cornell Digital Quality Index (DQI, http://www.library.cornell.edu/preservation/tutorial/conversion/conversion-04.html), developed from guidelines for the micrographics industry, though currently restricted to "printed text, line art and book illustrations," can be used for bitonal and grayscale scans, by measuring the smallest meaningful area of the image and using a QI to calculate the optimum dpi.

The UK's Higher Education Digitization Service (HEDS) has a Matrix for Potential Cost Factors (http://heds.herts.ac.uk/resources/matrix.html) which can be useful when considering different resolutions. Another good resource to refer to is the California State Library Scanning Standards (http://www.library.ca.gov/assets/acrobat/scandocrev1122.PDF) but you should use such instruments with great care.

The main principle to remember is that, once the resolution has been identified for a particular batch or collection, it should be fixed for those materials for the duration of the project.

Method of tone reproduction

When you undertake the actual scanning of your images, you will need to choose between several methods of tone reproduction. This phrase refers to the number of bits

(or “bit depth”) sampled for each pixel, and to the technique used to capture the full range of tones (or density) in the source materials.

Scanners record tonal values in digital images in one of three general ways: black and white, grayscale, and color. In black-and-white image capture, each pixel in the digital image is represented as either black or white (on or off). You can choose the threshold at which a given point will be considered to be black, and you can use halftoning algorithms to create a “screened” pictorial image, but black and white scanning is generally appropriate only for digitizing text and line art. In 8-bit grayscale capture, the tonal values in the original are recorded with a much larger palette that includes not only black and white, but also 254 intermediate shades of gray. Again, the thresholds can be controlled—either manually or automatically—as a first step to ensure that meaningful information in the highlight or shadow areas of images are adequately captured, and to get optimal results for originals with different degrees of contrast and overall darkness. In 24-bit color scanning, the tonal values in the original are reproduced from combinations of red, green, and blue (RGB) with palettes representing up to 16.7 million colors.

Your decision about which method to use for tone reproduction begins with the choice of black-and-white, grayscale, or color scanning as the default minimum for various classes of material. Your user needs will be an important determinant, but you should not assume that there is a simple correlation between increasing information capture and increasing user benefit. If your original is a faded black-and-white photographic print, for example, color scanning might accurately reproduce the effects of aging. However, file sizes for the color images would be three times larger than their grayscale counterparts. If color is not needed to meet the stated viewing, printing, and publication objectives for your project, grayscale may be the better option since it will be faster to download and smaller to store. Relatively speaking, adjusting resolution is a very straightforward process. Adjusting the settings for tone reproduction, however, can become complicated. Assume that each incremental jump might require changes to equipment, to operator skill, or both.

Having selected bit depth, the most important decisions regarding tone reproduction relate to reproducing the densities of tone in the source materials. Good digital masters are characterized by having pixels distributed across the tonal range in the image from black to white. Stated in the negative, tones are not “clipped” in good masters. Following scanning, all subsequent image processing relies upon pixels either being in the region of interest or not. A scanner that automatically chooses to reproduce the middle range of tones in an image might not have enough dynamic range to capture the meaningful highlights and shadows at both sides of the spectrum. Because experts say that much of the secret to good color reproduction is getting the neutrals right, the method used to capture densities should be carefully planned. For reflective materials such as photographic prints, grayscales might be used to set the “aimpoints” for black, white and middle gray in the digital masters. With transmissive materials, however, these decisions might need to be made by the operator on a case-by-case basis, which is one of the reasons that it can take much longer to scan negatives than prints in order to produce pleasing reproductions.

In addition to user needs, you should also consider which method will actually capture the information of your originals best. Grayscale imaging is appropriate for continuous tone images which lack color. Black-and-white imaging might be warranted for line art if cost savings are a concern; for instance, a black-and-white scan of a typescript page might provide adequate legibility. However, many forms of line art (for instance, manuscript pages) contain subtle color variation which might be of importance in distinguishing overlays of pen strokes and identifying the order in which the marks were made. For these purposes, full color scanning would offer significant benefits.

Projects will occasionally benchmark their imaging by establishing a minimum resolution for the collection as a whole. Most often the resolution depends upon the original, the resources available and the use that you intend to make of the material. Sample testing provides a good way to benchmark sampling rate. Setting minimum requirements is generally accepted as the best practice for establishing the capture requirements, but this must be weighed against the resources available.

Projects with very long-term digitization efforts may have to address the changes in standards for creating digital resources. Improvements in scanning technology and the diminishing cost of storage have radically increased the cultural heritage community's minimum standards for capture quality, and digital objects captured ten years ago may have been scanned at what would now be an almost unacceptably low level. A museum or library with a large number of legacy images captured at lower resolution will need to decide how to handle the transition. One possible strategy would be to digitize new material at the current recommended standard, and create from these new scans a set of lower-resolution derivatives to use with the legacy materials, if internal consistency is a crucial consideration. Over time, it may be possible to rescan the legacy materials and slowly upgrade the entire collection. Under most circumstances, it would be a waste of time to scan the new materials at an arbitrarily low resolution. The conversion of material from analog to digital will always be done in an environment of advancing technology and improving standards. You should be sure that you strike a balance between the costs, preservation suitability of the format, and the best usable standards. Generally, best practice is to capture at as high a level of overall quality as is appropriate to handling, quality, budget, or use considerations in order to limit the upgrading that might be required otherwise.

The advice box below gives summary guidelines for selecting the mode of tone reproduction.

Practical Advice Box:

Some practical guidelines for selecting the appropriate scanning method

- Use black-and-white (1-bit) scans if the image is black and white and contains no shading.
- Use grayscale (8-bit) for material with shades of gray or continuous tones: black-and-white photographs, half-tone illustrations, typescript, archival materials in black and white where, say, ink density varies and paper tonality appears to be important.
- Use 24-bit color or higher whenever color is meaningfully present in the original.[1]

Some projects scan at higher than 24 bit-depths (32-bit color 4 color channels of 8 bits each captured as red, green, blue, and 8 bits of grayscale or 48-bit color) even though the current generation of applications cannot render this depth at the moment. This is a hedge against future improvements in technology that you might wish to adopt. For originals of wide color gamut—e.g. paintings or natural history specimens—it may be important to select capture devices that allow the storage of raw files of 12, 14 or 16 bits per pixel, if an extremely high degree of color accuracy is a goal. Dynamic range may be checked and the image archived for future output devices that will be supported. Document the specific characteristics of the capture device, e.g.. non-interpolated highest resolution, spectral response captured IT-8 targets with the 50% gray neutralized, etc.

Transmissive material (for instance, slides and negatives) requires special treatment when scanning. The decisions made for resolution and bit depth are the same as for reflective material, but as slides and negatives tend to be smaller in size the resolution must be higher to achieve a digital image of useful size and quality. Although film scanners have generally been recommended to capture transmissive material, today many mid- to high-end flatbed scanners outperform production slide scanners. Projects at the Library of Congress capture 35mm slides up to 3500 dpi (the range is from 1200 dpi to 3500 dpi). Individual projects will identify the resolution required depending on the original source, the detail of content and the users' requirements.

You also need to ensure that there is a match between the dynamic range of the material that you are scanning and the abilities of your equipment to capture that range. Dynamic range is expressed as a scale from 0.0 (perfect white) to 4.0 (perfect black). Some color scanners miss subtle differences between dark and light colors and as a result do not create faithful digital representations of the content. The better quality the scanner, the greater the dynamic range. Drum scanners can capture between 3.0 and 3.8 and give good color quality; however they are very expensive and impractical for large projects. The scanner's ability to capture dynamic range is an important factor when deciding which equipment to purchase and this is especially true if you are working with transmissive material. Do not believe the manufacturer's claims of dynamic range. Instead, make your own tests by scanning a standard 22 step Kodak film gray scale and subtracting the highest from lowest perceptible densities. A scanner might have a high dynamic range but also introduce unacceptable noise in the dark areas of the image. If you cannot do

your own testing, you may be able to find another institution that has recorded results and is willing to share them.

Knowing your sources will allow you to ensure that you have configured your equipment appropriately. For instance, halftone images are made of regular patterns of dots or lines (the word “halftone” is derived from the printing technology method of representing photographs or artwork). They can be either color or grayscale. During digitization, the scanner dot sample combined with the halftone dots can produce undesirable patterns known as the ‘moiré effect’. One way of avoiding this is to scan at high resolution, but you may also need image enhancement processing, either post process or at scan time. Most good quality capture software packages will enable descreening to minimize this kind of scanning artifacts, although you may have to do this post-processing, using an image processing tool such as PhotoShop.

Process and Management Issues

Choosing a file format

In choosing the appropriate scanning method you must not only identify the appropriate digitization standard but also the best file format. You should select file formats that are platform-independent and supported by a wide variety of applications, both for browsing and editing. It is generally understood that for master images, uncompressed TIFF (Tagged Image File Format) provides the most suitable preservation format. ‘Uncompressed’ means that all the information encoded during the scanning process is retained in full. Digital projects store their master archived files as TIFF and then create smaller or derivative files for display and access for the users. These derivatives may be created using a number of different formats.

The file formats typically used for display and access are JPEG (which will gradually be displaced by JPEG2000), GIF, MrSid, PNG, and PDF. These file formats use compression algorithms to reduce the size of the digital files. Some algorithms, such as the LZW (Lempel-Zif-Welch) encoding scheme used by GIF do this to some degree without disposing of information altogether. LZW, for example, uses a lossless algorithm to compress files by up to 33% without throwing away data irrevocably. When the file is decompressed it is possible to reconstruct the compressed data. Other approaches to compression, including fractal and wavelet compression, take more radical approaches and sacrifice data to minimize file sizes. JPEG (Joint Photographics Experts Group) takes advantage of ‘lossy’ compression algorithms, and in particular fractal compression.MrSid (Multi-Resolution Seamless Image Database) uses wavelet-based image compression, which is especially well-suited for the distribution of very large images. The Library of Congress uses MrSid to deliver maps from its collections. As well as having impressive compression capabilities with limited visible information loss, it stores multiple resolutions of images in a single file and allows viewers to select the resolution (in pixels) that they feel will be most appropriate. There are other file formats that are suitable for delivery of images in different environments and you should consider

which of these will be most appropriate for your project. The table below provides an indication of the most commonly used formats.

Definition Box:

There is a range of non-proprietary and proprietary image file formats available. This table includes some of the more common formats for raster images.

Extension	Meaning	Description	Strengths/weaknesses
.tiff, .tif	TIFF (Tagged Image File Format)	Uncompressed file. Originally developed for desktop publishing. 1 to 64 bit depth. Used mostly for high quality imaging and archival storage.	Generally non-compressed, high quality. Large file sizes. Most TIFF readers only read a maximum of 24-bit color. Delivery over web is hampered by file sizes. Although LZW compression can reduce these file sizes by 33% it should not be used for archival masters.
.gif	GIF (Graphics Interchange Format)	This 8-bit file format has support for LZW compression, interlacing and transparency.	Lossless compression. Popular delivery format on web. .png was defined to replace GIF.
.jpg, .jpeg	JPEG (Joint Photographic Experts Group)	Compressed images. 8-24 bit. Variable amount of compression to vary quality and file size.	Lossy compression. Widely used delivery format. Flexible.
MrSid	Multiresolution Seamless Image Database	image-compression technology	Lossy compression. can compress pictures at higher ratios than JPEG; stores multiple resolutions of images in a single file and allows the viewer to select the resolution.
.pcd	ImagePac, PhotoCD	Lossy compression. 24 bit depth. Has 5 layered image resolutions.	Used mainly for delivery of high quality images on CD.
.png	PNG (Portable Network Graphics)	Lossless compression. 24 bit. Replaced GIF due to copyright issues on the LZW compression. Supports interlacing, transparency, gamma.	Some programs cannot read it.
.pdf	PDF (Portable Document Format)	4-64 bit depth. Uncompressed. Used mainly to image documents for delivery.	Need plug-in or adobe application to view.
.pct	PICT	Compressed. Mac standard. Up to 32 bit. (CMYK not used at 32 bit.)	Supported by Macs and a highly limited number of PC applications.

File sizes

You will need to determine how much storage space your image files will require and this is best done by working out the file sizes first. There are different approaches for scanners and linear scanning digital cameras. In the Calculating Box, Estimating Approximate File Sizes, we provide you with an example of the method. The quality of the master images will reflect the work processes you put in place, the equipment you use, and the resolution (dpi) and bit-depth (BD) you select. External variables that may limit this are the amount of storage available and the time for scanning each document.

Calculation Box:

Estimating Approximate File Sizes (in bytes):

Approximate file sizes for material created with a flat-bed scanner can be determined using the following formula:

$FS = (SH \times SW \times BD \times dpi^2)/8$

FS = file size

SH = Source Height (inches)

SW = Source Width (inches)

BD = bit depth

dpi = dots per inch

/8 because 8 bits = 1 byte, the unit in which file sizes are measured.

Large file sizes make compression essential for delivery to users. Derivative files in JPEG format are interoperable within the community and accessible through common access systems, i.e. the Internet. In some cases a much smaller thumbnail version of the image is used for quicker access (for instance, in displaying search results), allowing the user to download a larger file once they have found the image for which they are looking.

Calibration

Calibration is a crucial aspect of the quality control process in any digitization project. You will need to calibrate the work environment, the monitor, the capture devices, and the light sources. Frequently during the life of the project you will find it necessary to recalibrate this equipment as well.

First, establish the physical environment where the digitization work will be carried out. Avoid fluorescent lighting, natural lighting, reflections, and bright colors on the walls. Ideally, darken the space with black out curtains or dark neutral walls, to eliminate

ambient lighting, so that the only light source is the scanner bulb or the cold lights used for the digital camera. Where possible, maintain a dust-free environment with operators wearing neutral colored clothing and protective shoe wear. Avoiding carpets on the floor reduces dust accumulation. While it may not be possible to meet all these conditions, it is crucial to ensure the absence of ambient lighting — do not set the equipment up in a sunny room with fluorescent lighting! Consult conservators to insure that the temperature and humidity conditions will be stable and appropriate for the original work.

Once you have established the physical space, calibrate the monitor of the workstation at the beginning of every day, to ensure that the contrast, brightness and gamma settings are consistent. As monitors improve, gamma settings should remain constant, but do not assume that this is the case. A simple calibration program found in image processing software such as PhotoShop can help to do this calibration. Establish the settings for contrast and brightness at the beginning of the project within the adapted space and make sure all operators adhere to them. A set of guidelines with the appropriate settings will ensure the consistency of the digital images. Data about the settings of the capture device should appear as part of the metadata associated with the digital file. For more critical work you should use a spectrophotometer and calibration software such as the Eye-One Monitor (see http://www.gretagmacbeth.com).

Calibrate the equipment used for the digital process regularly. Most flatbed and film scanners can only be calibrated by their manufacturer because calibration is done at hardware rather than software levels. High-end digital cameras may have to be calibrated as the focal length and lighting may change with daily use. We recommend that calibration settings be defined at the start of each batch. Whether the work is being done in-house or outsourced, an appropriate policy or contract arrangements should be in place to ensure equipment is regularly recalibrated.

Color management

A Color Management System (CMS) should be employed by any institution wishing to accurately reproduce color from the source material (original object, artifact or film master) throughout the entire chain of digital hardware devices including monitors, dye-sub and ink-jet printers, film recorders and printing presses. A CMS is a group of software tools and hardware measurement instruments that work together to map the wide gamut of the original color space into the narrower gamut of display and paper output so that a reasonable and perceptible consistency and quality is maintained.

These systems are complex, evolving, and beyond the scope of this *Guide*. However, there is a growing need, both in terms of economics and image quality, for institutions to develop expertise and deploy a CMS. While lithographers, pre-press experts, image capture specialists and web designers all may use some form of a CMS they often are grounded in different nomenclatures. Therefore a standard form of communication is required to achieve the desired results in image distribution.

Fruitfully, there is a growing body of literature on the subject. The lingua franca is currently found in Adobe Photoshop version 7, which leverages Apple's Colorsync and Windows ICM operating system core color technologies to provide cross platform color management. These systems use "ICC profiles" that are based on rules made by the International Color Consortium for storing color information.

The most accessible writing on the subject has been by Bruce Fraser who is preparing a mass market reference on color management for autumn 2002 release.[2]

Link Box:

Color Management Web References:

International Color Consortium (ICC): http://www.color.org/

Apple Corporation's ColorSync: http://www.apple.com/colorsync/

X-Rite: http://www.xrite.com

Munsell Color Science Laboratory: http://www.cis.rit.edu/mcsl/

Electronics for Imaging: http://www.efi.com/

GretagMacbeth: http://www.gretagmacbeth.com/

Targets

Targets provide a mechanism for benchmarking the capture process. In simple terms, a target is a sample with known characteristics that can be used to establish a baseline or standard by which to assess capture. Your project should adopt a policy on what targets and scales to use, when to use them, and how their quality will be controlled. Two types of targets are commonly used: edge and detail (resolution) targets and color or grayscale charts. See Resolution Targets Example Box in Section VIII: Quality Control and Assurance.

Ensure that the targets used are appropriate to the material being scanned, as they are different for transparencies, prints, and certain kinds of reflective material. You should use new targets at the start of your project as they lose their accuracy with age. Scan the appropriate target at least once a day, at the start of a new batch of material, or if the settings of the equipment are changed for any reason. Although some think that targets should be attached to every image, this is not possible for non-reflective formats and will have resource implications as images then have to be cropped for display (though it is sometimes possible to automate this process). Including a scale with each digital image can be of value. If it is not in each image it may be sufficient to include it in the first image of a batch and whenever the settings of the equipment are adjusted (e.g. the position of the camera).

Metadata

Metadata associated with digital objects is the key to their sustainability. Metadata assists in the management of the digital data, documents its origin and specifications, and assists with discovery and retrieval. For detailed information on metadata, see the Appendix on Metadata. Briefly, from among the many types of metadata there are three that are generally used with digital images:

- Descriptive metadata describes and identifies information resources, and is used at the local level to enable searching and retrieval and at the web level to enable users to discover resources. Examples of descriptive metadata include Library of Congress Subject Headings, Categories for the Description of Works of Art (CDWA) (http://www.getty.edu/research/institute/standards/cdwa/), the Art and Architecture Thesaurus (AAT) (http://www.getty.edu/research/tools/vocabulary/aat/), and the Dublin Core metadata set. The title and creator of the object are examples of descriptive metadata.

- Structural metadata captures information about the structure of a digital object and the relationships between its parts, for instance between page images and text transcriptions. It facilitates navigation and presentation. Examples of structural metadata are included in the METS standard (http://www.loc.gov/standards/mets/) and SMIL (Synchronized Multimedia Integration Language) (http://www.w3.org/TR/REC-smil/).

- Administrative metadata facilitates short and long term management and processing, records data on creation and quality control, rights management and preservation. See Cornell University's "Web Hub for Developing Administrative Metadata for Electronic Resource Management" (http://www.library.cornell.edu/cts/elicensestudy/home.html). Another very useful resource is the NISO IMG draft standard on metadata requirements for digital still images: http://www.niso.org/standards/dsftu.html and http://www.niso.org/standards/resources/Z39_87_trial_use.pdf

A number of issues arise when designing a metadata model for a project and these must be addressed at the outset if you are to ensure consistency:

- Static and evolving elements: will the metadata item need to be changed frequently, or is it constant for the whole project?

- Manual and automatic creation: many systems have automatic metadata capture (for technical elements) but some metadata may have to be created by the operator during the capture process. You will need to assess the importance of this manually created metadata and its role in your collection and delivery plans. Keeping manual metadata creation to a minimum will help reduce costs. But for

some projects, detailed metadata created by experts will enhance the project's value so greatly that the cost will be more than justified.

- Who captures the metadata? The operator or an administrator? Very often this task is split between two or more people, particular if subject expertise is required.

Projects, on the whole, use a mixture of automatic capture and manual capture by the operator, subject specialist, or cataloger. The level and depth of metadata captured vary from project to project according to resource levels, including both staffing and equipment.

The operator of the digital camera or scanner will usually capture those technical metadata (as part of the administrative metadata set) that are not automatically captured by the scanning software. Descriptive metadata, particularly those relating to the original object, should be prepared in advance by a staff member with specialist knowledge of the collection.

File naming conventions

Before beginning a project you should define a file naming protocol so that your filenames will be consistent and intelligible. The filename can be part of the metadata and the information in the filename can be used to reflect the collection details, e.g. collection name, book title, or page number. One of the most common pieces of metadata in file names is an ID number. In file names you should always ensure that the three letters after the period are reserved for the file type information (.tif, .jpg) as this is important metadata for the operating system to know which applications can handle this file type.

Images of Text

Although the capturing and encoding of text have been dealt with in the previous section, here we will raise some of the issues in capturing pages of text as images, rather than as encoded text.

As already indicated, text can be captured in three ways: as an image only; as fully transcribed text; or as fully transcribed text with an associated image. There are currently no viable technologies for performing text searches on images, so if searchable text is required, some sort of transcription will be a necessity. However, for many purposes the original presentation of the text is also important, in which case an image of the text may be a useful addition. A source that is visual in nature as well as textual, for instance, an artist's book composed of both image and text, is best represented as both data types.

Example Box:

Projects that link images and text include:

- MOA (Making of America), using MOA2/METS: http://moa.umdl.umich.edu/
- Electronic Text Center at the University of Virginia Library: http://etext.lib.virginia.edu/
- American Memory at the Library of Congress: http://memory.loc.gov/
- Online Archive of California: http://www.oac.cdlib.org/
- AMICO: http://www.amico.org/

Projects that have only captured images of text because of the condition of the original material (for example, manuscripts or very early printed books), often store the best quality images so that when OCR technology improves sufficiently, or when more resources become available, they can create machine-readable text from these images.

Although proprietary file formats are generally not recommended for digitization projects, especially for archival versions, the PDF (Portable Document Format) can be very useful for a deliverable page image. You will have to buy the software to create PDF files (Adobe Acrobat), but the software to view PDF files is provided free (Adobe Acrobat Reader). The ability to create PDF files from any application and PDF's platform-independence make it particularly useful for delivering text material over the web.

Definition Box:

Portable Document Format (PDF)

- A proprietary format created by Adobe, but licensed to other software developers.
- Combines a precise page image (including fonts, graphics, and links) with a searchable full-text version of the document.
- Content can also be saved in RTF (Rich Text Format) for re-purposing.
- Users can read PDF files using the free Adobe Acrobat Reader, but creating PDF files requires the purchase of Adobe Acrobat.

Post-creation

Quality control and assurance is an integral element of the creation process, and the best digital surrogates should be created to minimize the work involved (see Section VIII on Quality Control below). Standards, technology, staffing and financial resources are all part of this process. Of all of these factors, perhaps the most significant is operator skill and training; investment in both of these areas will pay off significantly in the long run. However, there are also post processing techniques that may be needed to enhance the digital image. This is a practice carried out by a number of projects and generally in batch mode. The standard program used is Adobe PhotoShop© that has various filters and tools to assist in post image processing. Tasks commonly performed after the capture stage include unsharpen mask, gamma correction, noise reduction, deskewing, and cropping. Although you can do a lot in an image manipulation package, it is no substitute for care and attention to best practice at capture time. It is more efficient to set the guidelines for capture and make sure the operators adhere to these than to run image processing to redress errors made in the capture process.

The dynamic range (or density range) of a scanner is a secondary, but important consideration. If a scanner has a high dynamic range, it can sense a wide range of light levels. Dynamic range is usually higher in drum scanners and slide/film scanners. The dynamic range does not actually indicate how many different levels can be resolved but determines the smoothness of transitions between adjacent tones.

How scanner and camera manufactures implement this technology varies. A table in the Appendix on Equipment indicates the types and uses of scanning equipment available.

Conclusion

The successful digitization of still images begins with a careful examination of a set of complex issues. From selection of materials to the best means of capture and storage, these issues must be carefully considered for each project. Questions the project manager asks may include the following. Which materials will be selected for scanning—originals or derivatives? If fragile originals are digitized, what safeguards should be in place to protect them? What file formats are most appropriate for the material? What are appropriate file sizes and what will be the impact on storage requirements? What is the best way to develop a file-naming protocol? What about sustainability? This section has answered many of these questions or provided a framework within which to begin to seek answers. For further insight, the reader should consult Section VIII on Quality Control and Assurance. That section further clarifies procedures that should yield a quality product.

[1] That said, it is worth bearing in mind that Manfred Thaller's work at Duderstadt (http://www.archive.geschichte.mpg.de/duderstadt/dud-e.htm) and work commissioned by the Library of Congress (http://www.loc.gov/) concluded that for manuscripts except under special circumstances grayscale scans provided sufficient information.

[2] Bruce Fraser, Fred Bunting, and Chris Murphy, Real World Color Management (Peachpit Press: forthcoming, autumn 2002). 600pp; ISBN: 0201773406.

[3] For digital cameras this excluded consumer digital cameras that tend to use Contact Image Sensors (CIS) for weight and cost considerations.

[4] This description relates to standard (Legal sized, 8" x 14") flatbed scanners. Outsize (or large format) flatbed scanners are also available and employ similar technology. Sheet feed scanners employ an optional attachment to a standard scanner to enable automatic batch scanning of loose-leaf material.

[5] Engineering Scanners are also referred to as Wide Format Scanners.

VII. Audio/Video Capture and Management

Introduction

Despite differences in format and standards, the fundamental issues for capture, storage and management of audio and video are quite similar and will therefore be considered together. The interviews conducted with audio and video digitization projects highlighted two broad issues.

The first is concerned with terminology and definition: when considering audio and moving image material, what exactly is meant by digitization? For example, the copying of fragile, nitrate-based filmstock to a digital tape format such as digibeta format after restoration, or the audio transfer of a 78 rpm or wax cylinder onto DAT (Digital Audio Tape) is, strictly speaking, a digitization activity. In the context of this *Guide* digitization implies the analog-to-digital conversion of audio and video materials and their encoding in digital audio and video file formats that can be stored, manipulated and delivered using a variety of software and media (e.g. CDs and the Web). In this digital form, audio and video materials can be used and distributed in a richer variety of ways than in their analog form.

Secondly, the usual factors that come into play when considering capture standards and storage or delivery options for any type of digital object (such as the nature and conditions of the originals, the purpose of digitization, the intended mode of access, the needs and expectations of the intended audience) are the same for audio and video digitization projects. However, the fact that high-quality storage is still difficult and expensive, as well as the very high commercial stakes involved in networked delivery of high-quality audio and video to home users, all put a rather different slant on these familiar issues.

One further consideration of time-based media that does not apply to still image media is the act of editing for content. For instance, in a still image of an art artifact, one would normally not apply much "content-editing" to the master file, perhaps only cropping out the color-bar to create a sub-master for automated creation of publicly deliverable derivatives. However, editing a piece of audio or video for content, cutting out pauses or entire sections of the video that will not be used, is a necessary prerequisite for creating any kind of deliverable user version, and represents a large investment of time which needs to be taken into account when creating masters. It is important to keep one master copy of the “raw source” material that has not been edited for content, as a permanent reference file. But it can also be very useful to keep one version of the file as a sub-master that has received basic content editing, but is not technically different from the master file. Saving this sub-master file will save you the work of redoing this basic

content editing for each derivative, and allow for quicker, automated generation of edited derivatives. Saving this edited sub-master will have some associated costs in storage and management, but will save the even more expensive cost of re-editing later on.

This section looks at the motives for digitization and why sample projects have chosen particular digitization strategies, briefly describes the decision points and principles of audio and video capture, outlines the main standards and formats and their suitability for different access scenarios, and considers options for delivery and management.

Why digitize?

In common with other types of digitized media, digitized audio and video data should be easy to handle and manipulate and more convenient to deliver than their analog counterparts. By digitizing analog materials, we unlock the content from a fragile storage and delivery format, and make it possible for the content to be copied without loss.

Digitization facilitates research and study, allowing quick comparison, searching and editing within the digital object. In their digital form, audio and video content can be more effectively accessed by users than has been possible with analog collections. Once data is in digital form it can be converted more easily to another digital format without loss of quality, unlike all analog formats, which degrade with each use and lose quality when copied (an extreme example of this is audio wax cylinders which have very limited playing life, but every playing even of a vinyl record contributes to its destruction). Digitization is used, for example in the National Library of Norway, to preserve fragile and vulnerable materials (e.g., volatile nitrate-based film stock) or materials which need special handling or obsolete playback devices. The challenge here is to produce a high quality digital version. It is very time-consuming to quality check digital audio against the original analog material.

The downsides are financial (e.g., considerable investment in equipment, and large storage is necessary if high-quality masters are to be stored), technical (e.g., methods of compression are still evolving, high-bandwidth networks are not yet universally in place), the difficulty of data recovery from digital tapes in comparison with analog formats, and the continuing uncertainty about the suitability of digital formats for preservation. In digitizing video the cost of matching the quality to that of the original remains a formidable challenge. This is hard enough with tape/video sources, and is still very expensive with film sources. The Library of Congress still specifies analog audio tapes as preservation copies; the Survivors of the SHOAH Visual History Foundation specifies digital Betacam tape copies as the main preservation medium for film. The National Library of Norway argues that digital video formats are not yet good enough, and storage system resources are insufficient in size to make feasible the extensive reformatting of analog material into digital form. Of course, the main problem is that it is very expensive to create a digital version of analog film or video material of comparable quality, though the price of creating accurate digital copies of video, especially VHS, is currently much less than achieving the relative accuracy in copying film. It is common practice among film archives, such as the British Film Institute (http://www.bfi.org.uk), to create analog

copies, known as sub-masters, of their tape and film masters for viewing and exhibition purposes. A digitized version of the tape is just another, and better, way of making a viewing/exhibition copy for all the reasons outlined above.

Access to analog playback devices

Institutions may find themselves with a rich array of materials in analog form, but without the devices to play this material back. Unlike textual and still image material (with the exception of slides and born digital), audio and moving image material require a playback device in addition to a digital capture device. For example, a flatbed scanner can digitize directly a wide range of reflective media of different formats and sizes (e.g., photographs, letters, printed matter, bus tickets). No similar general-purpose capture device for audio and moving image material exists. A collection that included 78 rpm records, compact cassettes, 8 mm film and VHS video cassettes would require a playback device for each of these and each would then need to be connected to an appropriate digital capture device. For audio and moving image material that is already in a digital format (such as CD or Digibeta), playback equipment is less of a problem. Although many—frequently incompatible—proprietary digital formats exist, their recent development means suitable playback equipment is still on the market and relatively easy to source. Therefore this section concentrates on identifying analog audio and moving image formats, their properties and the source device required.

Three methods can be used to progress from moving image film to digital. Film can be transferred onto videotape for digitization via a transfer box or multiplexer. Both these options depend upon the material being projected in some way. Transfer boxes project the image into a box containing a mirror and onto a rear image projection screen with the video camera mounted on the other side of the box. The resulting video is subsequently digitized. These transfer boxes are not expensive, but do not in general produce as high a quality material because they produce generational loss in quality.

A better solution is to use a multiplexer. In this device the projector and camera are mounted on a single table. The image is projected by a set of lens and mirrors, directly into the camera without the need for a projection screen. This has advantages for image clarity. In both processes quality suffers because it introduces an extra production generation into the reformatting of the analog material. An alternative to these methods is the use of 8, 16 and 35 mm film for a chain film scanner to digitize directly from the analog film material. These machines scan the films and digitize at the scanner, passing the digital signal to the computer. (They work slightly differently for digital video. In this instance they grab individual lines of video to construct a frame and produce broadcast quality digital video.) In 2001 the costs of these machines remains high at between $500,000 and $1,000,000. One of the strengths of chain scanning is that, because the analog to digital conversion is done at the camera rather than on the computer, there is less opportunity for noise to be added by the process to the analog signal. Whereas small institutions can probably set up a transfer box or multiplexer system, even wealthy

institutions would find outsourcing to a facilities house to be the only practical option if they wished to go directly from the analog film to the digital material.

Determining a film's original frame rate is also difficult without viewing the film with a projector, particularly for old 8 and 16 mm films. The widespread availability of VHS and S-VHS video players makes the playback of these video formats for digitization relatively simple. The rapid adoption of digital formats in broadcasting, post-production and amateur markets is making the availability of even quite recent analog video devices scarce.

As there are fewer analog audio formats, these provide less of a problem than moving images. Compact cassette players, 33 and 45 rpm record players are still widely available new. Even record players with a 78 rpm speed can still be purchased new. The other formats present a greater challenge. If digitizing the sound as played on period equipment is important, the tone arms of phonographs and gramophones can be customized to provide an appropriate feed. Alternatively, the sound can be recorded via an external microphone onto a more convenient intermediary format. Reel to reel tape, wire recorders and cartridges pose similar problems of transfer. By modifying the equipment, it may be possible to provide a direct sound output. Alternatively, the sound can again be captured via an external microphone to an appropriate intermediate format. Here is where a great deal of specialist advice can be helpful. Just as we noted that it is easier to train a good photographer in digitization than it is to train a digital expert in photographic principles and practice, you will find that sound engineers bring to the digital environment strengths that are difficult to replicate.

In the case of all audio and moving image material, whether it is in analog or digital form, projects should carefully consider the advantages of outsourcing digitization. In general audio and moving image digitization require more and more expensive and specialized equipment than is necessary for still image material.

Audio Media	Properties	Source Device
Wax or Celluloid Cylinders	1890s & 1900s, up to 5”diameter, 2-4 mins. playing time	Phonograph. See http://www.tinfoil.com for details of digital transfer.
Wire	Magnetic coated wire drums or reels. Invented 1898. Widely used by the US military in WWII. Eclipsed by magnetic tape by the mid 1950s.	Wire Recorder
78 rpm shellac resin discs	1898 to late 1950s, 10” (25cm) and 12” (30cm) most common sizes	Gramophone (wind-up) or Hi-Fi. Gramophone’s steel needles need replacing after each side or record played. Hi-Fi needs a 78 rpm turntable and a cartridge with a 78 rpm stylus. For best results on modern equipment a phono pre-amplifier is required to correctly equalize the different types of record.
45 rpm and 33 rpm vinyl discs	7” (20cm) single and 12” long play (30cm). Long play (LPs) introduced in 1948, stereo recordings in 1958.	Hi-Fi. Hi-Fi requires turntable with 45 and 33 rpm speeds.
Reel to Reel magnetic tape	1/2” to 1/4” magnetic tape. BASF and AEG developed 6.5mm ferric tape and Magnetophone player in Germany from 1935. Post-war development in USA by Ampex and 3M. Stereo capability from 1949.	Reel to Reel player for appropriate width of tape.
Compact Cassette	Magnetic polyester tape introduced by Philips in 1963.	Hi-Fi. Hi-Fi requires compact cassette player.
Cartridge	1/4” magnetic tape. Fidelipac (4-track, devised 1956, released 1962) and Lear (8-track, 1965) cartridge systems.	Despite similarities 4 and 8 track cartridges are not compatible and require separate players. Predominantly used for in-car audio. 4 track unpopular outside of California and Florida.

Decision points for audio and video capture

There is a series of decisions to make in digitizing audio and video materials, having to do with hardware and software components, sampling rate and precision. To understand these decisions clearly, it may help to first explain the principles of analog and digital recording, and the digitization process itself.

In analog audio recording, a plucked string (for example) vibrates the air around it. These airwaves in turn vibrate a small membrane in a microphone and the membrane translates

those vibrations into fluctuating electronic voltages. During recording to tape, these voltages charge magnetic particles on the tape, which when played back will duplicate the original voltages, and hence the original sound. Recording moving images works similarly, except that instead of air vibrating a membrane, fluctuating light strikes an electronic receptor that changes those fluctuations into voltages.

Sound pressure waveforms and other analog signals vary continuously; they change from instant to instant, and as they change between two values, they go through all the values in between. Analog recordings represent real world sounds and images that have been translated into continually changing electronic voltages. Digital recording converts the analog wave into a stream of numbers and records the numbers instead of the wave. The conversion to digital is achieved using a device called an analog-to-digital converter (ADC). To play back the music, the stream of numbers is converted back to an analog wave by a digital-to-analog converter (DAC). The result is a recording with very high fidelity (very high similarity between the original signal and the reproduced signal) and perfect reproduction (the recording sounds the same every single time you play it no matter how many times you play it).

When a sound wave is sampled using an analog-to-digital converter, two variables must be controlled. The first is the sampling rate. This rate controls how many samples of sound are taken per second. The second is the sampling precision. This precision controls how many different gradations (quantization levels) are possible when taking the sample. The sampling error or quantization error means the fidelity of the reproduced wave is not as accurate as the analog original, basically the difference between the analog signal and the closet sample value is known as quantization error. This error is reduced, by increasing both the sampling rate and the precision. As the sampling rate and quantization levels increase, so does perceived sound quality.

In digital representation, the same varying voltages are sampled or measured at a specific rate, (e.g. 48,000 times a second or 48 kHz). The sample value is a number equal to the signal amplitude at the sampling instant. The frequency response of the digital audio file is exactly half the sampling rate (Nyquist Theorem). Because of sampling, a digital signal is segmented into steps that define the overall frequency response of the signal. A signal sampled at 48 kHz has a wider frequency response than one sampled at 44.1 kHz. These samples are represented by bits (0’s and 1’s) which can be processed and recorded. The more bits a sample contains, the better the picture or sound quality (e.g. 10-bit is better than 8-bit). A good digital signal will have a high number of samples (e.g. a high sampling rate) and a high number of bits (quantizing). Digital to digital processing is lossless and produces perfect copies or clones, because it is the bits that are copied rather than the analog voltages. High bit-depth is also result in much-increased dynamic range and lower quantization noise.

Ideally, each sampled amplitude value must exactly equal the true signal amplitude at the sampling instant. ADCs do not achieve this level of perfection. Normally, a fixed number of bits (binary digits) is used to represent a sample value. Therefore, the infinite set of values possible in the analog signal is not available for the samples. In fact, if there are R

bits in each sample, exactly 2R sample values are possible. For high-fidelity applications, such as archival copies of analog recordings, 24 bits per sample, or a so-called 24-bit resolution, should be used. The difference between the analog signal and the closest sample value is known as quantization error. Since it can be regarded as noise added to an otherwise perfect sample value, it is also often called quantization noise. 24-bit digital audio has negligible amounts of quantization noise.

With this background established, we can return to the practical questions a digitization project must ask. The first decision to make in digitizing audio materials involves hardware and software components. Digital audio can be created either by recording directly to the digital sound card or by using an external device to transfer audio material. High quality external devices produce superior results to sound cards; for archival digitization purposes, a high quality stand-alone ADC is recommended. Most internal PCI audio cards are built from inferior quality components and are prone to electrostatic interference from the computer circuitry.

Sample values over times are most commonly encoded in the PCM (pulse code modulation) format. This is the foundation of the digital audio file. PCM data can then be transmitted via a number of digital interfaces (such as AES/EBU) to other devices or software applications.

The next important decision to be made when making an analog to digital audio or video transfer, for example from vinyl or audio cassette or VHS videotape, is on the sampling and bit rates — in other words, the quality of resolution at which the transfer is to be made. Different sampling rates have an important effect on the end result. Nothing can compensate for a bad decision at the sampling stage, so the decision has to be informed, including careful consideration of purpose, intended longevity, circumstances and needs. Put plainly, the higher the number of samples, the better the resulting quality. Current technology allows audio digitization at the so-called “DVD standard” (96,000 Hz/24 bit) and should be recommended as the preferred audio digitization standard for most organizations. However, the quality of the original also needs to be taken into account: there is no point in using a high sampling rate for a poor quality original.

Related to the decision on sampling rate is the purpose of the digital transfer and the intended target audience and mode of delivery (e.g., is a preservation master at the highest possible quality necessary? Are users to access the materials via slow home Internet connections?). Of course deciding at what rate to sample has time, labor, and cost implications. Will it be possible, and cost-effective, to re-digitize the original source material at a later date for another purpose? Will the analog material be accessible in the future? Are they so fragile that you only have one opportunity to digitize from them? Are the access devices becoming increasingly rare? If not, then a better quality initial digitization is recommended to ensure cost-effective future uses. As we have noted elsewhere in this *Guide*, once material has been reformatted it is rare that the work will be done again. It is thus usually better practice to capture at the highest rate you can afford, and deliver a downsampled version, than to capture at a low rate now simply

because your immediate intention is to provide the material as small video images over modems on the web.

A policy decision has to be made on whether to clean up or enhance recordings and this, again, depends on the purpose of the digitization: is the objective to restore the recording, to re-create the sounds and images that reached the original recording device, or to make an accurate re-recording of the original? Filtering and noise reduction techniques that remove audio hiss, clicks and pops in an old recording inevitably change that recording and cut out some of the original sound.

Different organizations take different views according to their objectives. The Library of Congress, for example, takes a conservative stance on noise suppression for the preservation masters of historical recordings, seeking to reproduce the original as a recording before cleaning up or enhancing copies for specific listening or study purposes later on. Similarly, for much folk audio material it is important to provide a faithful digital representation of the original archival material, and even if, for example, there is a dog barking in the background of the singer's performance, it should be left in. From this master version it would be possible to produce derivatives in which the barking dog was removed if you wished to provide one group of listeners with access just to the singer's performance for easy listening and to produce a copy of the master with all the details of context in terms of environment and capture device (e.g. lack of noise suppression) for folk historians and anthropologists.

Given the flexibility of the digital audio file, it is recommended to digitize at the highest available settings (e.g., 96 kHz/24 bit) without any outboard or software digital signal processing (DSP) applied. The only exception may be a high-quality adjustable compressor/limiter to help with really noisy and soft signals. All other DSP techniques can be easily applied in the post-production process, and their choice should be determined by the delivery purpose, mode, and the target audience.

There may be exceptions to this general rule. For example, if making a digital transfer from an audio cassette it may be appropriate to use Dolby to get the best possible re-recording from the original. Indiana University's Hoagy Carmichael collection equalizes 78 rpm recordings but not reel-to-reel tapes. Uses of such techniques vary according to needs and circumstances; professional studio practices and methods for adjustment and preparation of equipment (e.g. cleaning and demagnetizing tape heads, distortion filtering, alignment testing) may be beyond the resources of one digitization project but vital for another. Once again this should never be done to the master, but may be done to derivatives depending on the purpose you are trying to achieve with them.

One should not forget about calibration and adjusting equalization (EQ) curves. Some analog recordings will require the use of calibration and an appropriate EQ curve (e.g., many vinyl recordings) to approximate the signal characteristics intended by the original mastering engineer.

The choice of digitization standards should not be contingent upon the type of acoustic signal to be digitized. While it is true that speech does not have the same dynamic range as the sound of a symphony orchestra, this should not justify the use of a lower bit-depth for speech recordings. We should apply the same, high standards to all kinds of acoustic signals.

Standards and formats: audio

The use of standards increases the portability of digital information across hardware platforms, space, and time. There are in general two types of standards in the marketplace, those that are proprietary and those that are non-proprietary. Proprietary standards are frequently developed by a single company or consortium and are designed to provide that organization or group with market advantages. Non-proprietary ones may also be developed by commercial consortia or not-for-profit groups, but the architecture of the standard is publicly accessible and often in the public domain. Three main audio formats are in common use:

- Microsoft's WAVE (.wav) format has been for a time the *de facto* standard for high-quality audio capture and preservation masters on PCs, and has largely overtaken AIFF (.aif) format. Projects in a Macintosh environment are still using the AIFF format. Of the projects interviewed, Harvard University Library uses the AIFF format for capture and preservation of sound files as AIFF includes structural metadata in the form of channel definitions and time marks.

- MPEG 1 Layer 2, fast being superseded by MPEG 1, Layer 3 (MP3) format offers high quality sound that comes close to .wav quality but at greatly reduced file sizes, achieved by bit-rate "lossy" compression. MP3 files therefore download to users much more quickly than .wav files and compression reduces space needed for storage and network bandwidth.

- Streaming formats such as RealAudio (.ra) which allow listening "on the fly", as the sound reaches the user's computer, eliminating the need to download a complete file.

Definition Box:

Audio Formats:	Extension	Meaning	Description	Strengths/ weaknesses
	Liquid Audio Secure Download	Liquid Audio is an audio player and has it's own proprietary encoder. Similar to MP3 it compresses file for ease of delivery over the Internet. Only AAC CD encoder available.	Boasts CD quality. Compressed file, thus some loss.	
.aif, .aifc	Audio Interchange File Format	Developed by Apple, for storing high quality music. Non-compressed format. Cannot be streamed. Can usually be played without additional plug-ins. Allows specification of sampling rates and sizes.	.aifc is the same as aif except it has compressed samples.	High quality. Flexible format. Large file sizes.
.au, .snd	SUN Audio	Mostly found on Unix computers. Specifies an arbitrary sampling rate. Can contain 8, 16, 24 & 32 bit.	In comparison to other 8 bit samples it has a larger dynamic range. Slow decompression rates	
.mp3	MPEG-1 Layer -3	Compressed format. File files vary depending on sampling and bit rate. Can be streamed, but not recommended as it isn't the best format for this — RealAudio and Windows media are better.	Typical compression of 10:1. Samples at 32000, 44100 and 48000 Hz.	Small file sizes. Good quality.
.paf	PARIS (Professional Audio Recording Integrated System)	Used with the Ensoniq PARIS digital audio editing system. Can contain 8, 16 & 24 bit.		
.ra	Real Audio	One of the most common formats especially for web distribution. Compresses up to 10:1.	Sound quality is passable, but not high quality. Lossy compression.	
.sdii	Sound Designer II	Originally digital sampling and editing platform. The format is still in use. Used mostly on Macs by professionals. It's a widely accepted standard for transferring audio files between editing software.	Problems with playing on PCs. High quality. Large file sizes.	

.sf	IRCAM	Usually used by academic users. 8 or 16 bit, specifies an arbitrary sampling rate.		
.voc	Older format, .wav files are far more common. Used mostly in IBM machines. It samples in relation to an internal clock.	Is not a flexible format.		
.wav	Wave	Windows media non-compressed format. Can usually be played without additional plug-ins. Specifies an arbitrary sampling rate. 8, 16, & 32 bit.	High quality. Large file sizes. Can be used on both Macs and PCs	
MIDI	Musical Instrument Digital Interface	Good for instrumental music. The file play digitally stored samples of instruments which are located on a sound card.		

It may be useful to be able to make a simple comparison between the file sizes of three of the formats. For example, a five minute music file will be some 60MB if stored in .wav, 5MB as an MP3 file, and about 1MB as a RealAudio file.

The MPEG standards are among the most important for digital audio and video. The Moving Picture Experts Group (MPEG, http://www.cselt.it/mpeg/) develops standards for digital audio and video compression under the auspices of the International Organization for Standardization (ISO). Each of the MPEG standards is designed for a particular purpose and is continually being developed. It is most commonly encountered as a means of delivering compressed video over the World Wide Web but these standards have also made interactive video on CD-ROM and Digital Television possible. The commonly encountered audio format MP3 is in fact a version of the MPEG-1 audio layer 3 standard.

MPEG standards have progressed considerably and care needs to be taken when using the term “MPEG format” (see table *Development of MPEG Standards*). MPEG 1, 2 and 4 are standard formats for encoding audio-visual media, MPEG 7 is a metadata standard for describing audio-visual media while MPEG 21 is a descriptive framework to encompass the creation, delivery, use, generation and transactions of digital objects. Projects that are intending to encode audio-visual material should be aware that MPEG 1,2 and 4 essentially define the decompression standard: the technology at the user’s end that puts the compressed stream of data back together. It is individual companies that control the encoding technology that compresses the data to be sent. When MPEG 1 was introduced, technology companies such as Microsoft and Apple envisaged a utopian future and included decoders in their software. When MPEG 2 was introduced the likes of

Microsoft, Apple and Real Networks decided the cost of MPEG 2 decoding licenses was too high and enhanced their existing technology. These provide high-quality, but proprietary AV streams supported by the distribution of free players (decoders for users). These systems can encode MPEG 2 but distributing MPEG 2 encoded files is problematic because it has been overtaken by proprietary formats such as Real. Therefore, for most projects seeking to encode AV material in an MPEG format, it is MPEG 1 that is a realistic option.

Definition Box:

The Development of MPEG Standards:

MPEG Format	Properties
MPEG 1: Started in 1988 and released in 1992. A standard for the storage and retrieval of moving pictures and associated audio on storage media	Designed for coding progressive video at a transmission rate of about 1.5 million bits per second. It was designed specifically for Video-CD and CD-I media. MPEG-1 audio layer-3 (MP3) has also evolved from early MPEG work.
MPEG 2. Started in 1990 and released in 1994. A standard for digital television.	Designed for coding interlaced images at transmission rates above 4 million bits per second. MPEG-2 is used for digital TV broadcast and digital versatile disk (DVD). An MPEG-2 player can handle MPEG-1 data as well.
MPEG 3. Merged with MPEG 2 in 1992.	A proposed MPEG-3 standard, intended for High Definition TV (HDTV), was merged with the MPEG-2 standard when it became apparent that the MPEG-2 standard met the HDTV requirements.
MPEG 4. Started in 1993, with version 1 released in 1998 and version 2 in 1999. A standard for multimedia applications that is currently being extended.	Designed to meet the convergence of telecommunications, computer and TV/Film industries and provide for flexible representation of audio-visual material.
MPEG 7. Started in 1997 and parts 1-6 (out of 7) released in 2001. A metadata standard for describing multimedia content data.	Designed to support some degree of interpretation of multimedia content's meaning by a device or computer by as wide a range of applications as possible.
MPEG 21. Started in 2000. A framework that is capable of supporting the delivery and use of all content types across the entire multimedia development chain.	Designed to provide a framework for the all-electronic creation, production, delivery and trade of content. Within the framework the other MPEG standards can be used where appropriate.

Sampling rates

As noted above, ideal sampling and bit rates depend on the nature of the original, but they are increasing as the technology allows. Simply put, sampling rate refers to the interval between points at which data are collected and bit-depth to the number of samples taken at any one sampling point. The comparison between digital audio and digital imaging is probably obvious; audio sampling rate (say 44.1kHz) is analogous to the number of pixels per inch (ppi) captured from a digital image (say 300 ppi) and in both cases the bit-depth relates to the number of samples taken at each interval point (say 16-bit stereo for audio or 24-bit color for images). Until recently a standard high-quality sampling rate was the CD-quality equivalent: 44.1kHz, 16-bit stereo; indeed this is the quality at which the Variations Project at the Indiana University (Bloomington) uses for capture and preservation. However, 48 kHz 16-bit is the sampling rate routinely used by the National Library of Norway for old recordings such as wax cylinders, and where the originals are of better quality 24-bit is used.

Very limited quality audio originals such as those in the Library of Congress's Edison collection were created from DAT tape at 22 kHz, 16-bit, mono. However, depending on the characteristics of the source item, the Library of Congress specifies 96 or 48 kHz as a sampling frequency for a master file as a future replacement for reel-to-reel analog tape recordings currently designated as preservation masters. In 2001, the American Folklife Center at the Library of Congress hosted a national meeting to discuss best practices for audio digitization. The consensus of the meeting was to move to 96/24 for preservation/archival purposes. We see no reason, given the declining cost of storage space, not to recommend 96/24 as best practice. Harvard University Library uses a sampling rate of 88.2 kHz for capturing and preserving and a bit rate of 24 for capturing and preserving and 16 for delivering. The file sizes created at these sampling rates are approximately 30 MB per minute at capturing stage and 1MB per minute at delivery. Research shows a clear difference in moving from 16 to 24 bit depth.

Standards and formats: video

Definition Box:

Moving Image Media[1]	Properties	Source Device
8mm & Super 8 Film	18 fps (frames per second) most common frame rate, followed by 12 fps and 24 fps (often used with sound film). The latter frame rate tended to be used by professionals or for capturing moving objects. During the early 1960s 18 fps started to appear. 8mm sound film appeared around 1960. Super 8 introduced by Kodak in 1965. It is perforated in the center and not the edges of the frame. 3” diameter reels are most common, 6” and 7” reels and cassettes are also found.	An 8mm film projector, for "standard" 8 mm and/or Super 8 film. Determining the original frame rate can be problematic. Most older projectors are variable speed which is useful. Projectors should be in excellent condition and the film unshrunken. Capstan drives are better for the film and the sprockets.
16mm Film	Very common film format	16mm film projector
35mm Film	Very common film format	35mm film projector
1/4” Reel to Reel Video Tape	Can be confused with audio tape. 10” reels are audio, some video, as well as audio, formats used 7” and 5” reels.	1/4” videotape recorder.
1/2” (12.5mm) Reel-to-Reel Video Tape		1/2” videotape recorder. Machine maintenance and replacement parts very difficult.
3/4” (U-Matic) Tape or Cassette	Broadcast TV format. U-Matic has been around since the early 1970s and remains a popular production and archive format because of relatively low cost compared to Betacam SP.	3/4” U-Matic machine. Come in fixed or portable, reel or cassette versions
1” Reel to Reel Video Tape		1” Reel to Reel tape player.
2” Reel to Reel Video Tape	Television programs from the late 1950s to 1970s.	2” Reel to Reel tape player. Playback equipment for this has become increasingly rare.
8mm Video Cassette	8mm video comes in two formats 8 mm and Hi-8 (equivalent to VHS and S-VHS)	Hi-8 players can play standard 8 mm cassettes but not vice versa.
1/2” (12.5mm) Video Tape Cassette	Betacam SP is a popular field and post-production format. M-II is a popular broadcast quality format. S-VHS is a higher quality format of the ubiquitous VHS home video cassette. The now obsolete Beta and Video 2000 formats also used 1/2” tape cassettes.	Betacam SP and M-II require compatible players. S-VHS players will play standard VHS cassettes but not vice versa. Although almost identical, S-VHS cassettes have additional holes in the casing.[2]

Three main file formats are in common use: MPEG (see table), QuickTime and RealVideo. However, both the Library of Congress and the National Library of Norway have held back from keeping preservation copies of film material as files on servers, but

rather have kept digital video preservation master copies on Digibeta tape. The Library of Congress uses the sampling ratio of 4:2:2 for digital tape copies, which is the current component digital tape recording standard. 4:2:2 refers to the sampling ratio of the three parts of a component color difference signal (one luminance channel and two chroma channels). For every 4 samples of the luminance channel there are 2 samples for each of the chroma channels. As usual, as the sampling rate increases, so the quality increases. In 4:4:4, the chroma channels are sampled equally to the luminance channel, creating better color definition, but this high sampling rate cannot easily be recorded onto tape.

Of the file formats that projects might use for service, rather than preservation copies, the highest quality films are likely to be stored in the .mpg (MPEG) format. The average file size for the MPEG 1 format is about 9 MB for each minute of film. The Library of Congress's MPEG 1 files are created at 30 frames per second at a data rate of approximately 1.2 Mbits per second of playing time. The National Library of Norway makes digital transfer from film copies in MPEG 1 at 1.5 mbits per second, at a resolution of 25 frames per second, or MPEG2 at from 6 to 15 mbit per second.

QuickTime may include a variety of compression methods; some higher end, some lower end. For instance, QuickTime (with Cinepak compression) offers smaller, downloadable files and allows films to be viewed on lower-end computer systems. The Library of Congress's QuickTime files are created at 10-15 frames per second at a data rate of approximately 640 Kbits per second, usually quoted as 80 Kbytes/sec of playing time. The average file size in the QuickTime (Cinepak) format is about 5 MB for each minute of motion picture. The Berkeley Art Museum/Pacific Film Archive (BAM/PFA) currently captures video content as DV (see DV discussion). This DV stream is then converted and saved as a video master file in QuickTime/DV format. Derivative files are extracted at a much smaller resolution and are delivered online in QuickTime/Sorensen format for video, and QuickTime/Qualcomm for audio-only materials. Content digitized so far by BAM/PFA includes videos of artist talks and works of video-art from the collection. Works on film will require a different methodology.

RealVideo is a streaming format allowing viewing of the moving image material as it arrives at the user's computer and thus eliminating the need to download the file completely before viewing. Real Media format is especially useful for computers with slower Internet connections, such as a 28.8kps modem. Video playback is slower (3-6 frames per second), may be affected by Internet traffic, and currently provides an image of lesser quality than the worst broadcast TV. But it does make the distribution of material to wide audiences possible.

Definition Box:

Moving Image formats

Extension	Meaning	Description	Strengths/weaknesses
.mpg	Moving Picture Experts Group	Standards created by the group working for ISO/IEC. MPEG-1: for Video CD and MP3 are based on this early standard. MPEG-2: DVD based on this. MPEG-4: Standard for multimedia on the web. MPEG-7: Currently under development; for 'Multimedia Content Description Interface'.	Good quality and low file sizes. MPEG-1 can take a while to load.
.qt, .mov	QuickTime	Created initially for Macs, can now be used on PCs too. QuickTime player. Quick Time 4 has streaming capabilities.	Excellent quality, easy capture, widely used, can be large. In Windows the QuickTime player takes up lots of space.
.viv	Vivo	No updates since 1997. Played on VivoActive player. Video stream always sent over http (unlike Real Video or Windows Media). Bought by Real networks in 1998.	High compression rates, poor quality due to compression to maximize streaming, various incompatibility issues.
.avi	Audio/Video Interleave	QuickView, Windows' Media Player. Replaced largely by MPEG and Windows media.	Large files, very good quality, must be encoded/decoded properly,
.rma	RealMedia	Streaming format. Proprietary format that is an equivalent to Windows Media.	Requires RealMedia plug-in.
.wma	Windows Media Video	Streaming format. Version 8 offers near DVD performance.	

Not all audio and video material will need to be digitized from analog material; much of it will come, increasingly, from digital materials. In discussing still image material we noted that derivatives of digitized images should be measured in pixel dimensions rather than dots per inch so digital video formats are measured in pixels. Digital video in NTSC format consists of 720 x 480 pixels[3]. This is the standard resolution used in MPEG-2-compressed commercially distributed DVD movies. As you examine the chart on digital video formats, it will be obvious that the main differences between the DV formats relate

to the formats of the tapes (e.g. size and running time) themselves, but there are, in the case of DVCPRO 50 and Digital-S, some differences in the compression algorithm used.

Definition Box:

Digital Video Formats:

Format	Tape Size	Compressor	Compression Ratio	YUV sampling	Running Time (mins.)
DVCAM	6mm	DV25	5:1	4:1:1 NTSC 4:2:0 PAL	184
DVCPRO	6mm	DV25	5:1	4:1:1	183
DVCPRO 50	6mm	DV50	3.1:1	4:2:2	90
Digital S	12.5mm	DV50	3.1:1	4:2:2	124
Digital Betacam	12.5mm	Sony	3:1	4:2:2	94

Post-capture processing

Following the capture of digital AV material, significant post-processing may be necessary or desirable, either to clean up the data using digital processing techniques (such as noise reduction), or to produce smaller, downsampled and compressed versions for various kinds of distribution. Some of these processes can be automated, but they are still labor-intensive and need to be planned and provided for in advance. Examples of post-capture processing carried out at the University of Virginia Robertson Media Center include adding data, fades, and altering frame rate and size using Media Cleaner, Final Cut Pro or Imovie tools. Other projects may choose to clean up noise or edit moving image materials into easily accessible chunks. Post-processing should be performed on digital derivatives, with an unedited digital master copy kept for reference and archival purposes. It is worth considering the kinds of uses to which these derivatives will be put so as to sequence the kinds of processing you perform and minimize the number of special-purpose derivatives you must create and maintain. For instance, where initial capture is performed at a very high bit rate and sampling depth, as recommended in this section, some institutions produce two other masters to meet delivery needs: a de-noised "production master," and a downsampled de-noised master from which Real Audio files can be easily made using current software.

Audio-visual metadata

Metadata is a crucial extra element that must be created to accompany digital audio or video, and the principles of metadata interoperability and documentation standards are as important to digital AV media as to still image and text media. However, unlike textual

resources, audio and video cannot currently be adequately searched by themselves as a raw resource. There is no cheap, standard way to apply the equivalent of a "full-text search" to AV materials. As a result, metadata for audio and video is doubly crucial to internal management as well as public use of such resources. Metadata for digital audio and visual resources can be used in much the same way as metadata for complex digital objects composed of still images. For instance, a metadata standard like METS (Metadata Encoding and Transmission Standard) can be used to describe the structure of a digital object: for instance, a book that is represented by dozens of digital page images, plus a full-text transcription. METS allows one to connect an individual page image with structural information about that page (i.e. Chapter 10, page 99), as well as alternate representations of that page (say a transcription of the text). In this way, one can present and navigate the various digital files (page images, transcriptions, METS XML file) as one cohesive “digital object”. This allows users of the resource to search for phrases that may occur on a particular page, and be taken to the location of that page in the complex digital object. Similarly, AV metadata standards such as METS (with the appropriate extension schema), or others like SMIL (Synchronized Multimedia Integration Language) can be used to describe the content and structure of time-based digital files such as audio and video. SMIL, for instance, can be used to describe structural metadata about a particular frame of video (frame 30, timecode 01:20:36.01) as well as link the appropriate series of frames to alternate representations such as a transcription of the dialogue in that scene. As with image resources, this allows users to search for a particular bit of dialogue or the name of a character, and be taken directly to the video scene in which they appear. In addition to acting as a discovery tool, audio-visual resources metadata also helps enable the exchange of resources between institutions, and facilitates the internal management and preservation of such resources.

As of 2002, there is no shared good practice for what constitutes minimum metadata for digital audio and video in the cultural heritage sector. So, whether creating A/V metadata for management or for discovery, it is recommended to use or adapt a standards-based metadata standard like METS or SMIL, and to use the standard to capture as much information as the institution can afford. Note that some technical metadata can be automatically—and hence cheaply—generated; for instance, some cameras create and can export metadata about capture date, format, exposure, etc. Metadata about content is more expensive to create, since it requires human intervention. At a minimum, useful retrieval will require a basic metadata set for each file, based on a standard like the Dublin Core. In addition, any information that will require effort to rediscover later on (for instance, the format of the original material, or its provenance) should be captured immediately if possible. Most institutions will probably decide to create a very basic metadata set for each file, with the possibility of adding further information later in the lifespan of the digital object. Fewer institutions will be able to create a full metadata record for every object from the outset, although arguably this may be the more efficient way to proceed, since it consolidates processes such as quality checking.

Link Box:

Key AV Metadata Sites

- Dublin Core Metadata Implementers: http://www.fiu.edu/~diglib/DC/impPurpose.html
- Synchronized Multimedia Integration Language (SMIL) http://www.w3.org/AudioVideo/
- Metadata Encoding and Transmission (METS) Standard http://www.loc.gov/mets
- MPEG-7: http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm
- Authority Tools for Audio-Visual Catalogers: http://ublib.buffalo.edu/libraries/units/cts/olac/capc/authtools.html#g
- Authority Resources for Cataloging Popular Music: http://www.music.indiana.edu/tech_s/mla/wgpms/wgpms.htm
- Library of Congress's Digital Audio-Visual Preservation Prototyping Project: http://lcweb.loc.gov/rr/mopic/avprot/avlcdocs.html#md
- Library of Congress's Digital Audio-Visual Extensions to METS Standard http://www.loc.gov/rr/mopic/avprot/metsmenu2.html
- Cinemedia's SWIFT project for on-demand delivery of film and video: http://www.cinemedia.net/SWIFT/project.html

Options for delivery and management

The main issues to be examined when considering options for delivery, storage and management are connected with the larger file sizes associated with audio and moving image material. The restricted nature of available bandwidth for delivery and the lack of appropriate data storage facilities pose challenges for institutions. Rights protection is another important issue, in which streaming as a delivery option can be part of a solution. For example, in Indiana University's Variations project, while the material in copyright is digitized under the legal provision for libraries (preservation or fair use), students may view the streaming format but are unable to copy (and potentially misuse) it. At the Berkeley Art Museum / Pacific Film Archive, copyright issues have slowed efforts to digitize film and video content. Digitization has so far included documentary material and selected works of video art but no film, as of 2001.

While almost all of the project managers interviewed are committed to maintaining top quality archival masters, many are prepared at this time to trade off top quality against lower, compressed quality for ease of access and delivery. Similarly, they are prepared to trade material of longer duration (say full-length feature films) against availability of short (say 30-second) samples. These are purely pragmatic decisions, based on current technology and bandwidth.

Much depends on the purpose of the digitization project: if the objective is to provide ready access to audio visual materials (perhaps historical material), which are otherwise

very difficult to view or hear for a general audience, then delivery of lower quality, compressed versions fits the purpose. If, however, the aim is to preserve fragile materials, there is still some doubt as to the durability of most digital formats; most projects are keeping options open by storing preservation copies on tape until such time as server-based digital formats are proven for long-term preservation purposes.

For many projects, storage is problematic because of the massive data sizes of audio and video material. At the extreme ends of the spectrum, the Survivors of the SHOAH Visual History Foundation and Indiana University's Variations project serve as exemplars of large data storage options for video and audio material respectively. At Indiana, migration of MPEG files from the current tape storage system to a new university-wide mass storage system has recently been completed, with migration of WAV files currently in progress. The Digital Library Program is working with University Information Technology Services to ensure long-term preservation/access of Library objects in this mass storage system.

The system for storage and retrieval at the Survivors of the SHOAH Visual History Foundation involves use of local caches of 1 Terabyte for instant retrieval. If the information is not available on the cache then the disc server (180 Terabytes) at the Visual History Foundation is accessed. The final step is for the information to be found in the tape storage and uploaded to the server and local cache. This is done automatically by a system that uses a robotic arm to locate, pick, and load the desired information, but it can take between 5 and 10 minutes. This system is appropriate for a project with such a massive scope (52,000 video records of testimonies) and multi-million dollar budget, but much of its practice cannot be applied to smaller museums or archive projects as the level of funding required will not be available to produce such sophisticated systems and technology.

In summary, it is difficult to draw general conclusions or to make firm recommendations on ideal storage, delivery and management of digitized audio and moving image materials. Practice and technology are still evolving, although pointers to the future indicate that streamed delivery of highest quality audio and video material will become widespread, problems of massive storage will ease as costs decrease and reliable preservation formats will be found. Perhaps the most prudent course in the meantime is to transfer materials at the best sampling rate possible and store them at the best possible quality with as little use of lossy compression as the budget allows in order to keep the highest quality materials to work with in the future and eventually migrate as necessary.

---

[1] This table does not include details on esoteric formats such as lantern slides, picture discs, 4.75, 9.5, 17.5, 22, 28 and 70mm film formats.

[2] There are three different VHS formats: PAL which is common in Australia, New Zealand, the United Kingdom and most of Europe; NTSC used in the USA, Canada and Japan; and SECAM used in France, much of Eastern Europe and Russia. PAL (Phase

Alternate Line) uses 625 horizontal lines at a field rate of 50 fields per second (or 25 frames per second). Only 576 of these lines are used for picture information with the remaining 49 lines used for sync or holding additional information such as closed captioning. SECAM, (Sequential Couleur avec Memoire or sequential color with memory) uses the same bandwidth as PAL but transmits the color information sequentially. NTSC (National Television Standards Committee) is a black-and-white and color compatible 525-line system that scans a nominal 29.97 interlaced television picture frames per second.

[3] DV-PAL has a pixel dimension of 720 x 576.

VIII. Quality Control and Assurance

Introduction

This section discusses some of the most effective methods of Quality Control and Assurance (QC&A). In the fields of cultural heritage and humanities the authenticity, integrity, and reliability of digital material is crucial to users at all levels, whether scholars, teachers, students, or researchers. What is sometimes less obvious is that effective QC&A must be planned at the outset of the project and built into all its processes. For QC&A to ensure the quality of the project's deliverables this activity cannot be left until the end. If your QC&A procedures accomplish nothing except to reveal the errors that faulty processes have already produced, they are largely a failure.

At the outset it is useful to make the distinction between quality control (QC) and quality assurance (QA) as these terms are often interchanged. Quality control includes the procedures and practices that you put in place to ensure the consistency, integrity and reliability of the digitization process. Quality assurance refers to the procedures by which you check the quality of the final product.

For all classes of digital materials three measures of quality can be used: completeness, fidelity or faithfulness to the original, and legibility. Completeness simply measures whether the entire object has been captured, without cropping or other essential loss of material. Legibility is a functional measure which indicates whether the digitized version is intelligible: for text, whether the characters can be read; for images, whether the object depicted can be discerned at a basically acceptable level. Fidelity goes a step further and measures whether the digital version represents the original in a way that goes beyond simple legibility: a legible representation of a manuscript page will allow the text to be deciphered, while a faithful representation might also convey the visual texture of the paper and the gradations of ink color. Within these measures one can also make the distinction between subjective and objective measures. Comparing a digital image with the original using the naked eye is a subjective measure. Using a computer to log the number of potential errors per 1000 characters in OCR-generated text is an objective measure. In practice, both subjective and objective measures are combined in the whole quality control and assurance process. A digitizer evaluating color reproduction in a digital image might use a program such as Adobe Photoshop© to confirm that the RGB values for a color bar included in the image fall within the correct range. A computer may flag errors in a text but it is a human editor who checks the nature of these errors (if indeed they are errors) and decides whether to accept, reject or correct the text.

Related to this point is the need to be clear about the differences between tasks that are fully automated (e.g., comparing checksums to ensure success of data transfer), some that are semi-automated (correcting what software identifies as likely OCR-generated errors),

and finally, those that are fully manual (confirming that text has not been inadvertently cropped; confirming that image content is not skewed in the frame).

Finally, throughout this section it is important to be aware of the differences between the quality assessment of people and systems and the quality assessment of products. In the former case, you are intervening to assess the capacities of your staff and of the systems and products you use during the digitization and quality assurance process. For instance, you might administer a color blindness test to the staff who will be conducting the color correction and quality assurance of digital images, or you might use technical targets to ensure that a camera is operating consistently from batch to batch, or week to week. In the latter case, you are establishing systems to check the actual product that results from the digitization process. Examples of product quality assessment, which abound in the chapter, include using targets as a means to establish correct tone reproduction for the image products created by the photographer and digital camera, or checking a certain percentage of each batch of scanned images for skew and cropping as they come from the vendor.

Digital objects

QC&A in some form is an integral part of any digitization project, but the procedures that projects put in place are often informal, of variable consistency, unreliable, and under-resourced. The aim should be to embed QC&A in the project at the points where it will do the most good. The project must define which digital products, such as masters and deliverable images, metadata, encoded texts, transcriptions and translations, are to be included in its QC&A program. Once you have determined the range of the QC&A program, the next step is to establish appropriate QC benchmarks.

Although there are no broadly accepted standards for image quality, encoded text accuracy or audio quality, there are a number of generally accepted guidelines that can prove very useful. These will be discussed in more detail below, but in general your QC threshold needs to match the purpose of the digital deliverable. For instance, the QC requirements for digital images that will serve as replacements for deteriorating nitrate film would probably be much more stringent—given that there may never be another digitization opportunity—than for “access” images created from relatively stable black and white photographic prints.

In general your QC threshold needs to match the purpose of the digital deliverable.

The first step is to carry out an initial series of QC benchmarking tests on a sample of the analog materials to test capture settings and establish threshold guidelines for rejecting digital material that does not meet the quality criteria. This should be undertaken for each type of material to be digitized and for each of the different types of output that the

deliverable will take. For example, tests should be carried out with text for screen and print, or with audio for streaming or for preservation. In this way, different QC procedures, and possibly different methods of QA, will need to be established for different types of material. The resulting benchmark standards will then form the basis of an ongoing QC&A program. These QC benchmarks must be documented and presented to project staff in a way that makes them easy to implement and easy to monitor for effectiveness.

These QC benchmarks must be documented and presented to project staff in a way that makes them easy to implement and easy to monitor for effectiveness.

As well as defining your QC benchmark, you will have to decide on the scope of your QA program. Do you check the quality of every image or page of text against the original (100% check)? Do you undertake a stratified or random sample (for example, every 10th image, or a random 10%)? What QA procedure are you going to follow when a digital object is rejected? If a digital image is created from a surrogate will you compare the digital image against the original object or the surrogate for the purpose of QC&A? Also remember that even if your digitization has been outsourced, the project is still responsible for performing a QA check on the vendor's work, irrespective of how detailed the QC&A requirements are in the contract with the vendor. Images of England, for example, has put in place quality assurance procedures and a custom application that allows it to check the digital images, mark errors on screen and return the error details to the vendor to enable them quickly to identify and correct the error. Furthermore, will the digital images be compared against the originals or some other benchmark such as a color chart? At what magnification, if any, will the comparison take place?

It is also necessary to ensure that a project's QC&A program includes measures for controlling the digitization environment and the way staff work (e.g. ensuring they take regular breaks). If you have specified your hardware correctly (see Equipment in Section II on Resources, and in more detail in the Equipment Appendix) then you should have a system that is appropriate for the types of materials being digitized and the purpose for which they are being created.

In addition to ensuring the correctness and informational quality of the digital materials you are creating, you need to ensure the integrity of the data files themselves over the long term. This is a significant preservation issue, and is discussed further in Section XIV. But it is also an ongoing quality assurance concern and is worth touching on here. There are two points to be addressed: checking the integrity of files as they move through your workflow and are transferred from medium to medium, and checking the storage media at intervals to guard against failure and data loss. The first of these can be automated to a large extent. Checksums, for instance, provide a simple way of ascertaining whether a file has been altered or corrupted during data transfer. For SGML and XML documents, parsing and validation can help verify the document's integrity,

though they cannot detect every kind of error. Another more powerful option is a version control system, which can track all changes made to a set of files and prevent casual or accidental alteration. In a lengthy workflow process, where the files change hands and move around a great deal, basic steps like these can save you huge amounts of effort and trouble later on.

The second point—the quality of the media on which your digital objects are stored—is easily neglected. However, storage media vary greatly in their longevity and robustness, and you should not only assess which media are appropriate for your purposes but also take preventative measures to guard against loss. Although hard disk failure is relatively rare, running periodic scandisk checks and defragmenting drives on a monthly basis can go a long way to identifying bad sectors and preventing read/write errors as well as improving the performance of your computer in the process. Removable media such as CDs and tapes should be purchased from a reputable brand; batch diagnosis for structural validity of the media and files is a relatively efficient method of quality assurance. Certain media, such as floppy disks, JAZ cartridges, and ZIP disks, have relatively high failure rates which make them inappropriate for storage and backup.

There are a number of further steps that can be taken to minimize potential errors in QC&A for particular types of material, including images, OCR and encoded text.

Images

Image capture may serve a range of project goals, and for some of these only completeness and legibility are essential. However, for projects that require that the digital image be a faithful representation of the original (to the extent allowed by the medium), careful calibration of the equipment will be needed to ensure consistent and highly controlled results. You should also ensure that the material that you are to digitize is free of dirt and dust, that it is positioned on a calibrated scanner or camera, that the digital capture environment is adequately controlled (e.g. free from stray light sources), and that suitable control targets have been used (see box).

Example Box:

Resolution and color targets should be used.

Common resolution targets:

- AIIM Scanner Test Chart#2
- RIT Alphanumeric Resolution Test Object
- IEEE Std 167A.1995
- IEEE Standard Facsimile Test Chart

Common color targets:

- Q13 and Q14 Kodak Color Separation Guide and Grayscale Targets
- Q60 Kodak Color Target
- Kodak Grayscale Charts

Remember that color targets are made with organic dyes and that these dyes breakdown as they age. Therefore over time the charts lose their accuracy.

For all images there is a series of key QC&A tests to perform. The first set of checks is relatively straightforward. Check that the entire image has been captured (i.e. not cropped) including any captions or titles. Are pages missing or out of sequence? Is the image skewed? Does the image have the correct file name? The second set of checks is more complex to assess, and includes detail reproduction, tone reproduction, color reproduction, and color accuracy.

For images of textual material, line drawings, etchings, plans and other objects with distinct line-based features, detail reproduction is the key to image quality. When benchmarking, a resolution target or the smallest resolvable detail should be used. This provides a comparison point for examining legibility, completeness, sharpness, contrast, serifs and uniformity, paying particular attention to individual strokes and dense cross hatchings.

For grayscale and color images the bit depth and dynamic range are as important as resolution in assessing image quality. These issues have already been discussed in some depth in Section VI on images, and in the appendix on digital data capture. Briefly, bit depth is the amount of information (in bits) used to represent a pixel. The use of a grayscale or color chart can provide a standardized reference point for assessing the quality of color and tone reproduction. Assessing color and tone reproduction can be highly subjective, particularly if fidelity is desired, but features to look out for include the presence of details in shadows and highlights (an indication of a good dynamic range), and a smooth transition in tones, particularly on skin and sky (a blotchy or pixellated effect is an indication of insufficient bit-depth). Compare color, contrast and brightness to the original or to a color chart, paying particular attention if digitizing from negatives,

where simple inversion can produce a color cast, or digitizing from print, where a herringbone, or moiré, effect can be present.

Example Box:

What to look for when checking digital images for quality:

- image not the correct size
- image in wrong resolution
- image in wrong file format
- image in wrong mode or bit-depth (e.g. bitonal when it should be grayscale)
- overall light problems (e.g. too dark)
- loss of detail in highlights or shadows
- poor contrast
- uneven tone or flares
- missing scan lines or dropped-out pixels
- lack of sharpness
- excessive sharpening (including halos around dark edges)
- image in wrong orientation
- image not centered or skewed
- incomplete or cropped images
- excessive noise (see dark areas)
- misaligned color channels
- image processing and scanner artifacts (e.g. extraneous lines, noise, banding)

One of the best examples of an imaging QC&A system is that of the Genealogical Society of Utah (http://www.lds.org). Quality control and assurance of the images is an integrated part of the GSU capture system and uses software specially developed for the programs. The volunteers examine each image and reject for skew, readability and color balance. If rejected, the image will be recaptured and noted in the log file for re-indexing. When the images are sent to GSU from the projects, an audit program, again specially developed for the project, carries out further checks. A wizard sets up the rejection threshold and uses a random number generator to identify the individual items to be selected for the inclusion in the statistical sample. An auditor checks the image, and there are twenty-four possible rejection criteria; if an image is rejected then the reason for rejection is noted in the log file. If three images are rejected, the audit program turns off

and the whole batch must be re-imaged. Auditors are trained to use this system and to evaluate the images. Typically, they look at 150 images in a batch at their choice of speed, e.g. 1 sec per image. They also use histogram analysis as well as checksum, to facilitate automatic QC&A. (Checksum is a value computed from a block of data and transmitted and stored along with it to check whether errors have occurred in transmission or storage.)

The work environment has a significant impact on the QC&A of digital images.

You can control some QC&A environment factors with the following methods:

- Ensure the work area is free from strong colors and reflections (e.g. a room painted gray without direct sunlight)
- Always calibrate the monitor regularly (see Section VI on Images)
- Use appropriate software for your image or text format
- View originals and print output in bright, preferably natural, light
- View display output in low, but not dark, light
- Minimize the number of different people who perform QC&A measures, to reduce variability
- Ensure that all QC&A staff use identical equipment and configure it the same way
- Consider using color management software to ensure consistency throughout the digitization chain
- Document the QC&A procedure either through having online checklists or through paper-based forms that staff complete for each object

OCR and encoded text

When digitizing text, the page images are subject to the same QC&A checks as for line art images, but further checks are required at the OCR and encoding stages. The range, scope and method of QC&A must be established and appropriate benchmarks set. If you are creating OCR texts for indexing purposes or batch processing texts, then an error rate of 00.5% may be acceptable (this is the Library of Congress' NDLP threshold for batch processing by vendors), but if you are creating a scholarly edition, then nothing less than 100% accuracy may be required. OCR rarely produces results better than 99.9% accuracy, or one error in every 1,000 characters (roughly 10–12 lines). Double processing documents and checking each version against the other can speed up identification of errors, but frequently there is no substitute for manually checking the digital version against the original by second, or even third proofreaders.

The Thesaurus Musicarum Latinarum (TML http://www.music.indiana.edu/tml/) at Indiana University employs one such text QC&A system. The quality control procedures in place for the digital deliverables involve at least three sets of complete checks on each text. The individual entering the data is expected to check and correct the text before printing it out and passing it to the person responsible for proofreading and checking. This second person proofs the printout and marks any errors identified. As the corrections are made to the electronic version the marks on the printout are counter-marked. Then, both the printed text with its marks and counter-marks and the electronic text are passed to a third person for review, prior to approval and addition to the database. Where there is a high error rate at any stage in the process, the text is printed once again and subjected to a second proofreading, as outlined just above. The final check and approval by the project director facilitates consistency in the quality control process. Lessons they have learned from this experience are that it has been difficult to persuade people to proofread character-by-character (rather than word-by-word) and to refrain from global search-and-replace editing. In general, the TML has discovered that no more than four double-spaced pages of 12-point text can be proofread per hour with an acceptable rate of accuracy.

With XML or SGML encoded texts the use of a parser to validate files against a DTD greatly assists the QC&A process. Remember, however, that while the coding can be declared well formed and valid, a parser would not pick up typographical errors in the content of the text, or an incorrect choice of tag from a set of structurally valid choices. To catch the latter kind of error, more sophisticated (and usually project-specific) tools are sometimes developed which allow a reviewer to look at overall patterns of tag usage and spot anomalies without having to review every file in detail. For large-scale encoding projects such tools may repay the cost of development.

Audio and moving image materials

Digitizing audio and video materials in the cultural heritage community is a relatively new area of activity, QA in this area is especially new, and so there is little documentation of QA practices in libraries and museums regarding AV material. The feasibility of these methods for a community that must weigh scarce labor resources against the need to protect the investment in digitization is thus unfortunately not yet well established in practice. However, some significant examples do exist.

The Library of Congress ‘Audio-Visual Prototyping Project’ specifies the following QA in a contract document intended for vendors who are handling the digitizing of LOC audio materials:

“Contractor quality review of audio shall include, but is not limited to, the following criteria:

- complete item has been captured

- re-recordings are not flawed by noise or distortion beyond that present in the source recordings
- all files open and play properly
- re-recordings meet specifications for resolution, sampling frequency, and other formatting requirements
- recorded calibration tone sets meet requirements for reproduction quality"

Phase 2 of the European project, "Presto: preservation technology for audio and video" (http://presto.joanneum.ac.at/index.asp) includes the goal of automating AV quality control: "Implement automated quality control: audio and image analysis algorithm development, fitting of algorithms to the application, and cost / benefit analysis." Indeed, Presto partners are developing tools that can, for instance, track errors that occur in transferring film to digital and log the timecode of the frame of the film in which the error occurs so that humans may conduct special quality control measures on that area of the file.

While work in this area begins in individual institutions and projects, it is clear that there is no consensus in the cultural heritage sector about what constitutes good practice for quality control and quality assurance of digital audio and video files (metadata is covered elsewhere). Standard QA methods such as random selection of sample files for inspection and playback by a human, as well as other QA methods developed for digital images, are highly recommended as the very minimum while more robust and yet cost-effective means are being tested on a broad scale in the cultural heritage sector.

Metadata

QC&A of metadata is generally less well documented, but its accuracy is perhaps more important than that of the digital deliverables themselves. If users cannot find an object because of poorly checked metadata they will never know that you have meticulously checked and verified the audio, image, or text. There are even fewer guidelines for checking metadata quality than for images and texts, although for XML-encoded metadata the general principles for encoded text apply to a large extent. From among the projects in this *Guide* good practice indicates use of the following:

- An appropriate metadata standard (such as EAD, Dublin Core, TEI Headers, METS)
- Name authority files (for person and place names)
- Subject thesauri and classifications
- Data type controls on fields (e.g. text, number)

- Independent double checking of entries

Good practice projects have checked metadata at the time of image checking, but many projects have very little metadata QC&A. It is recognized that metadata QC&A is not done once, as with images, but is an ongoing process. As with any QC&A, this should be considered in the resources both at the creation of the project and for the future.

Conclusion

The QC&A techniques presented in this section have focused largely on detailed questions of accuracy and the scrutiny of the project data. However, you should be careful not to neglect QC&A in other areas of the project. Steering or advisory groups, project plans, and flow charts can all perform an important role in assuring the quality of project management. Similarly, documented procedures for the archiving and preservation of digital material, such as the frequency of back-ups, rewinding archive tapes, and moving copies off-site, all contribute to the overall QC&A environment of a project. Finally, be sure to take a global view of the entire project and the complete product it is creating: how do the parts work together? Is the overall design successful? Does it meet your users' needs? These larger evaluative questions require different methods from those sketched above, but are an important aspect of the overall assessment of the project's success.

IX. Working with Others

Introduction

As digitization moves from small, discrete projects, conducted within individual institutions, to larger, multi-departmental, multi-institutional and international digital programs, collaboration becomes an increasingly vital consideration. The funding, creation, access and sustainability of digital surrogates increasingly depend on arrangements in which institutions work with others. These may range from relatively simple cooperative arrangements to collaborations where partners develop a common goal and input resources jointly. They may also include working with consultants, paid or unpaid, who contribute specialized expertise but are not closely involved with project planning, or using outside agencies to accomplish specific segments of the project. This section of the *Guide* uses the term "collaborate" in the broad sense of any type of cooperative or partnership relationship.

Our interviews revealed that all the projects and programs that are establishing good practice in all areas of digitization already collaborate on various levels. As they improve their understanding of their collections and the digitization process, they increasingly recognize that collaboration has a crucial role to play. It quickly enables a team to widen its subject base, increase its available expertise, and leverage its equipment, human, and financial resources (see Section II on Project Planning).

Collaboration and cooperation have always played an important role in the cultural heritage sector, from international art exhibitions to simple inter-library loans for academic establishments. Fruitful use and understanding of our heritage depends on such collaboration. Digitization extends this rationale, and provides more efficient ways of improving access and increasing scholarship.

Any digital program starting out should explore all the possible layers of collaboration and also understand the advantages and possible difficulties that any collaboration can hold.

Why collaborate?

Institutions seeking to create digital versions of their collections may not consider collaborating as a necessary, or even desirable, option. Their aim may be quite limited and local in its scope: for instance, to digitize a particular section of a collection for a specific research purpose or a specific teaching task, where collaboration seems like an unnecessary complication. However, collaboration does not necessarily entail large-scale partnership over a number of years, but can be as simple as seeking advice and adopting standards that will streamline the current project and enable future collaborations. In

these cases, the advantages may lie in the greater breadth of experience applied to the project—an asset under any circumstances—or, more importantly, in ensuring that the local collection can interoperate effectively with related resources elsewhere. Collaboration may also help solve problems of permissions or digital rights management, by grouping together organizations with a common interest and giving them greater leverage when negotiating with rights-holders. For instance, Georgetown University encountered resistance from publishers when seeking permission to digitize critical editions of philosophical works. However, once the university joined with professional organizations such as the Hegel Society of America, they were able to negotiate successfully for digitization rights.

Collaboration does not necessarily entail large-scale partnership over a number of years, but can be as simple as seeking advice and adopting standards that will enable future collaborations.

Funding may also play an important role in encouraging collaboration. Seeking funding is a critical and time-consuming aspect of any digital project or program, from identifying likely funding sources to writing complex bids and complying with funders' guidelines and requirements. Sharing this work may be advantageous, particularly for smaller institutions. Increasingly, the major funders in Europe and in North America are seeking collaborative programs for successful digital projects. The purpose may be to exploit the full capacity of digitization, to unite geographically disparate collections and improve access, as well as to enhance partnership between institutions. It may be that collaboration is a necessary—or at least a highly desirable—aspect of any successfully funded program. The projects we interviewed all agreed that collaboration was absolutely necessary to achieve the depths of digitization that they desire for their collections.

Types of collaboration

There are many kinds of collaboration, ranging from simple consultancy to sharing skills or equipment, outsourcing, and fully-fledged multi-institutional international projects. Collaboration can take place locally, regionally, nationally and internationally. All these forms of collaboration require an understanding of the issues and careful planning and management.

Consultancy

Many projects seek the advice of consultants. One typical approach involves identifying a suitable company or person to visit the project, discuss the tasks involved, and produce a feasibility study or similar document on how the project should proceed. Consultants do not necessarily have to be external; in a large institution, there may be internal experts

who can offer advice on planning, creating and sustainability at the beginning of any project. Generally, it is a good idea to seek this internal advice before applying for funding to ensure that the bid is as complete and precise as possible. However, not all institutions have access internally to such expertise and may seek external advice. Remember that external advice will be costly and funders may have reservations about permitting feasibility studies unless they have been specified in the bid. It is worth exploring funders' attitudes to such costs in advance. Advice from an internal source may be negotiated in a more relaxed manner, perhaps with specific advice on funding bids and purchasing technical equipment. If you will need advice regularly from these internal sources throughout the project, then you may have to factor some sort of payment for recovery of salary costs. It is worthwhile exploring your institution to discover what expertise exists. It may also be possible to create a support group of similar projects that can work together, creating an institutional policy that can be applied to new projects, thus saving costs in time and staffing.

Any contract with an external consultant must be well prepared and managed throughout the process to ensure that the deliverables, such as a feasibility study, are produced on time and meet the specifications. This will take more time and cost more than you may anticipate, so think carefully about whether the costs of seeking professional external advice will actually achieve the goals you set. You may find you can reduce costs by using consultants who are located close to you, to reduce costs in travel and phone calls. However, one specialist notes that this could constrain one from getting the right person for the job, particularly if the expertise you seek is rare. A large-scale, multi-institutional project in the U.S. might serve its needs very well by bringing in an expert from the UK, for example, to help plan workflows or to train key staff. It is also advisable to seek references on any institution or company you plan to consult. Staff in universities, libraries and other institutions within the cultural heritage sector will be able to act as consultants or make recommendations. There are other avenues of expertise that projects can explore, such as this *Guide*, or other resources created by the community. Make use of the listservs and newsgroups to seek such advice. The Research Libraries Group (RLG) and NINCH have web pages that will help.

Joint projects — funding, resource sharing from local to international

The most likely collaboration is one that involves more than one project from more than one institution and then generally within the same sector. These projects use the expertise and skills from similar institutions or departments. The Making of America is one such project, involving the University of Michigan and Cornell University. The project describes this ongoing collaboration as follows:

"Drawing on the depth of primary materials at the Michigan and Cornell libraries, these two institutions are developing a thematically-related digital library documenting American social history from the antebellum period through reconstruction. At the University of Michigan, approximately 1,600 books and ten journals with imprints primarily between 1850 and 1877 were selected, scanned, and made available through the present system. Librarians, researchers, and instructors continue to work together to

determine the content of this digital library and to evaluate the impact of this resource on research and teaching at both institutions." (http://moa.umdl.umich.edu/about.html)

As a highly successful project that involved collaboration at all levels and at all stages of the digital process, The Making of America is worth further investigation if your institution is looking for a collaboration model to follow.

Collaboration in different shapes

The impetus for collaboration comes from a number of directions. Funding decisions are frequently made on the basis of the skills and resources available at an institution. Even large institutions may not have all of the skills and expertise required entirely in-house. Smaller institutions may only have their collections and specialist subject knowledge to offer. Even if institutions have sufficient capacity for digital content creation, the capacity required for effective delivery and sustainability may only be found at larger institutions or through pooling resources. Some digital projects will require a critical mass of content to begin with and others will require that their content be scalable in order to be sustainable. As digital projects give way to digital programs, particularly for delivery, the need to collaborate increases. Indeed, underpinning the widespread and convenient delivery of digital content is the need to adopt common standards and practices, which itself drives collaboration. In today's digital landscape the individual, entrepreneurial start-up has given way to the age of partnerships and mergers.

Underpinning the widespread and convenient delivery of digital content is the need to adopt common standards and practices, itself a driver of collaboration.

Digitization projects can collaborate in a wide variety of ways. Collaboration may involve the pooling of resources, or the sharing of content; it may operate at the institutional, local, regional, national or international levels; it may exist within or across sector boundaries; and it may involve formal or informal relationships. None of these types of collaboration in best in the abstract; different types will suit different projects and will affect the partners in different ways. Nevertheless, a number of trends are evident.

The library and museum fields offer fertile opportunities for cross-sector collaboration. The enhanced functionality that web delivery can provide, and users frequently expect, requires an increased amount of interpretative information. Museums and galleries are well versed in providing succinct information on the selected parts of their collections on display while libraries and archives have traditionally concentrated on providing users with access information to their entire holdings. Web delivery demands a combination of these elements: comprehensive descriptive information to facilitate resource discovery, and access and supplementary contextual information to aid interpretation. Cross sector

collaboration in physical resources, personnel, skills and experience can thus greatly benefit both parties.

> *For many smaller institutions some form of content collaboration may be the only method by which their collections can be digitized cost effectively.*

The ability of the web to deliver still images, text, moving images, audio and 3D representations offers another opportunity for digital content creators to provide new ways of accessing collections. Traditionally, researchers would access sources across physically dispersed collections that may be distinguished by the type, subject or provenance of the material held. Even if the digital representations of these collections remain distributed (and the distributed collaborative model is a common one) the opportunity to provide a unified point of access, and hence distribution, greatly enhances the functionality for users and the dissemination objectives of collection holders. Although the educational benefit of access to digital collections is often presumed rather than evaluated it is clear that a single point of access to interrelated and previously unobtainable resources makes the learning experience more efficient if nothing else. The institutional emphasis on education, particularly in the K-12 sector, makes such content collaboration a major consideration.

For many smaller institutions, some form of content collaboration may be the only method by which their collections can be digitized cost-effectively. The collaboration may take the form of licensing rights to a commercial organization in return for creating digital content, or providing access to collections to a larger non-profit institution. Even if projects can afford to undertake the actual digital conversion themselves, issues such as copyright and intellectual property rights, electronic access controls and effective delivery and asset management systems may be beyond their financial and personnel capabilities. Lastly, the ongoing delivery and preservation of digital content raises issues that lie beyond the scope of even the largest institution and require the collaboration of the whole cultural heritage sector if they are to be addressed effectively.

There seems to be no predominant pattern in the level at which collaboration should take place. However, smaller organizations do seem to require either a major institutional partner to act as a catalyst or a partnership with enough similar-sized organizations to gain critical mass. In both these cases a formal collaborative arrangement would be most appropriate. Collaborative groups of any size can bring in expertise less formally by including specialists on their advisory boards.

Outsourcing

Outsourcing is a growing method of creating digital resources. Projects or programs will identify a vendor (usually through a request for proposals or RFP), negotiate a contract

and give the vendor access to the material from digitization and to metadata creation, indexing and cataloging. This is a very simplistic view of the process; there are a number of factors that must be taken into consideration before approaching a vendor to ensure that both parties understand the process and agree on the outcomes. Outsourcing can be done on-site as well as off-site; in the former case the vendor will bring equipment and staff to the material. This may or may not be more costly, since the vendor avoids a whole range of infrastructure costs, but may have additional travel costs.

Vendors are increasing their knowledge of the types of documents in this sector and are now purchasing the specialist equipment to handle them. However, you should ensure that any vendor you are approaching understands the nature of your material. Scanning 10,000 business documents, on good quality laser print, is very different from handling 10,000 fourteenth-century church records on different types of material and of varying sizes. Show the vendor a sample of the more difficult material to ensure that they have the requisite expertise and appropriate equipment. You should also ask them to produce sample scans of some of your material to help you assess their abilities.

Make sure that the quality of work produced is of the standard required. This *Guide* has sections on standards for resolution, tonal range, color depth, and other variables (see Sections V, VI and VII). Quality control and assurance (QC&A) is a vital part of any agreement with a vendor (see Section VIII). The actual terms of the agreement will depend on cost and the vendor's capabilities. It is expensive to check every image, and depending upon the nature of material and how you are processing it checking a sample of the digital products may be sufficient, but do not neglect quality control and assurance as a cost saving measure. Agree with the vendor on a guaranteed level of quality control of the process(es) they will provide and require that they have quality assurance procedures in place. The only secure way for you to ensure that the agreed level of quality has been reached is to conduct a check on the output once it has been returned to your organization. For instance, check one in ten images and if any error (such as skew) is found in a specified batch, then that whole batch should be returned to the vendor to be digitized again. For obvious reasons, you should be sure to stipulate such quality assurance conditions in advance.

The most immediate benefit of using a vendor is that once the costs have been agreed, you are protected against unexpected cost increases as these would normally be carried by the vendor. This condition must be stated explicitly in the contract. Establishing a comprehensible and workable contract is an essential stage in the outsourcing process; if it is carefully worked out (using lawyers where possible), then outsourcing can be a great benefit to an institution.

Many projects in the survey reported that they had been reluctant to use outside vendors because of the nature of their material and a lack of confidence in the capability of commercial services to appreciate the conservation, handling, and content issues. Increasingly, though, commercial companies have realized that the cultural heritage sector has funds available for the creation of digital content. Firms with digital imaging

experience are investing in the expertise and facilities that allows the sector to entrust them with the digital process.

Nevertheless, outsourcing should be avoided when it requires rare and fragile material to be sent off-site. In these instances, a vendor should set up digitizing stations on-site so that the materials do not have to be moved. Generally, using vendors is most successful where the material is in good condition and of consistent type and quality. Variations in size and quality will cause the vendor as many problems as they would cause the project team, but vendors rarely have the content and conservation expertise to respond to these difficulties.

Do not underestimate the complexity of getting the analog material from storage to the vendor, having it returned to the institution after scanning, then checking it and returning it to storage. Catalogers and conservators should be involved in this process. The latter will be especially helpful if there are questions as to how material should be packaged, shipped or handled by the vendor. The operation should be planned, managed, and documented to suit the whole collection as well as the material to be digitized.

As increasing the volume of material to be digitized should reduce the per-unit costs, it is always worth asking at the planning stage whether extending the project would be sensible. Collaboration between projects within the same institution will also help reduce per-unit digitization costs by providing access to increased volumes of material.

Outsourcing is thus a viable option but one that requires building a solid relationship with a trusted vendor and managing every step to ensure the highest quality result.

Managing collaboration and partnership agreements

The success of any collaborative venture lies in its management, with firm aims and objectives set out for all partners. Goals and deadlines must be fixed, agreed and adhered to. Employing a project manager to take overall responsibility for ensuring that all partners deliver the work agreed is a distinct advantage. It is also worth ensuring that all partners are doing work that is appropriate to the level of skills and resources available to them. Establishing effective methods and mechanisms of communication between members of the consortium has always proved a challenge for collaborative initiatives. You should work to build good, well-documented channels for communication and decision-making, which involve all of the relevant participants.

All this should be set out clearly in a Partnership Agreement, and the more comprehensive it is, the more likely the project will be a success. The Agreement should make explicit the role(s) and responsibilities of each partner. It should differentiate between those activities that a partner will carry out in support of achieving their local project obligations and those that a partner must complete for the group as a whole (such as local training or dependable file access). The Agreement should be signed by director-level staff to ensure that ‘project buy-in’ has taken place at a senior-enough level in the organization.

Where you are working in a multi-institution collaboration, a consortium policy can make explicit the responsibilities of the central agent. The central agent is the institution that manages the initiative's financial resources; other sites involved are partners. It is crucial to examine which functions will be centralized and which will be de-centralized and thus under local control. For instance, will training take place locally or will it all be centralized? Will standards of QC&A and metadata creation be defined locally or centrally? Will the texts be stored on a central server for cross-resource searching? Will the digital images be stored centrally, or locally, or both? Some of these questions have obvious answers, but answers to others will vary depending upon the circumstances, material, and nature of the partnership. The Colorado Digitization Project (CDP), for example, has centralized training sessions for all partners and stores the texts and images on a centralized server that all partners can search; some partners also hold the images locally, with the CDP acting as a depository and collections manager. The MOAC project, based in Berkeley, stores all EAD and MOA2 XML documents centrally for searching by all the partners but these records link to images that are stored locally on each museum's server.

Further examples of the issues that must be addressed include the following:

- If the digital images are stored centrally, can the central agent make use of them for purposes other than those envisaged when the project was started?

- Can the central agent alter the context of the images without express permission from each partner?

- Who creates the upgrades to interface to the central resource?

These questions, and others, should be addressed at the start of any project. This will help forestall any problems or misunderstandings that might occur later in the process.

The Partnership Agreement applies to the funding and publicity stages of the project just as much as it does to the creation and usage of material. Issues to consider at this stage are:

- How much income will each site have, and who decides what percentage each site should devote to specific activities? For example, it may be that one site spends more on creation and another on interface design, depending upon the strengths of each partner.

- Does the central agent have a responsibility to funnel a certain average percentage of income to the partners for content development? Or will it hold on to funds for system development?

- Are the decentralized partners obliged to provide publicity or credit the central agency whenever publicizing the resource?

- Must the central agent always list the partners whenever publicizing the project?
- Must the partners participate in all fundraising efforts, or just some?

The Colorado Digitization Project (CDP) is an excellent example of how various partners can come together, combining their talents, resources and collections in such a way that all partners are satisfied with the management and organization of the collaboration.

Collaborations such as this enable large-scale digital programs to realize the full potential of digitization. The CDP brings together a variety of institutions from the state of Colorado with the aim of producing a digital resource that encapsulates material from the collections of museums and libraries throughout the state. A member of the CDP staff acts as the project manager, assisting and advising the projects, as well as setting minimum standards for creating images and metadata. Skills and resources are shared through training sessions and scanning laboratories located throughout Colorado. The laboratories provide a critical link in the chain because they house the technical equipment needed to enable the participation of projects from smaller institutions. The project enables both large public libraries with state-of-the-art equipment and small local museums with no equipment at all to collaborate in creating a high quality digital resource. Overall management of the digital resource is controlled from a central point, but this is crucial to the success of any collaborative project; no matter how large or small the contribution of the institution is, the goals and deliverables must be agreed upon to ensure the initiative's success.

A successful collaboration depends on establishing guidelines and creating standards that all partners adhere to. These guidelines will cover such areas as definition of types and depth of metadata, descriptive terminology or control vocabulary, file formats (whether for images, audio or video), standards for interface design, and guidelines for QC&A. This rule applies to all collaborations, whether between two departments in a single institution or as part of a multi-institutional, international project.

It is also important to apply long-term thinking to collaborations and to recognize that it may be beneficial to participate in more than one type of collaboration, so as to maximize each partnership. It is also advisable to think about the life-cycle of a collaboration and the conditions that might bring it to a graceful close. For instance, does the collaboration have an expiration date? If not, is there a date when all agreements are revisited? Has an exit strategy been devised for each partner as well as for the central agency? These questions have to be considered if an institution means to make the best of its collections and resources and ensure that they remain accessible in the future.

Successful collaboration can enable scattered collections to be reunited as well as producing materials that enhance and complement each other, providing rich digital resources for users.

Thinking Box:

Key factors to consider when starting a project:

- Investigate whether there are areas where collaboration could be possible
- Plan for the possibility of collaboration at the outset
- Look for strategic fit when selecting collaborators
- Commit to equal participation in shaping projects and sharing the benefits
- Involve all partners at all stages
- Establish clear communication lines and document communication
- Moderate and manage expectations, including those of donors and funders
- Adopt shared project goals and focus (for example, create a mission statement)
- Establish a steering group that reflects institutional types, experience, resources, and knowledge of the partners
- Do not amplify institutional hesitation — act, don't just talk
- Have flexibility to accommodate different aims, structures and needs of partners
- Document areas of responsibility, and track progress
- Ensure accountability — one person to enforce milestones and timelines
- Embed interoperability — minimum guidelines for flexibility
- Cement collaboration with face-to-face meetings with minutes and formal agreements
- Pool resources
- Avoid duplication of effort
- Trust, respect, and endeavor to understand professional skills of the other collaborators
- Think ahead about access and distribution to ensure equitability of access and benefits
- Ensure partners have compatible exit strategies

X. Distribution

Introduction

This section looks at how users gain access to your digital objects and how they use them. Once the data (text, images, sound, moving images) and metadata have been created there are two sets of decisions to be made: choosing appropriate delivery option(s) and modes of access. These are key decisions and are likely to be made at an early stage in the development cycle. Delivery options cover questions of how to distribute the data assets and grant users access to them; in other words they are concerned with dissemination, or 'publishing'. Access modes—including applications and interface design decisions—affect how users will be presented with the material and how they may search, retrieve, navigate and manipulate them.

Delivery options

Keep in mind that the delivery options discussed in this section pertain to delivery versions of data and to public versions of metadata. Information pertaining to the storage and distribution possibilities pertaining to *masters* is in Sections XIII-XIV. Essentially, there are two choices for distribution: using a network as the medium (principally the Internet); or using portable media (usually magnetic disks, optical discs, and tapes). We consider both choices, including a note of requirements for each, and then compare their advantages and disadvantages.

Portable and removable media

There are three categories of portable and removable media that projects may wish to use to deliver their digital content. These are: tapes (e.g. Digital Audio Tape/Digital Data Storage (DAT), Digital Linear Tape (DLT); optical discs (CD-ROM or DVD); and magnetic disks (floppy diskettes, Zip disks, JAZ drives). There are a number of issues to consider, including the robustness of these media, and in particular their relatively high sensitivity to magnetism, atmospheric conditions and temperature; cost; the level of hardware support among the user base; user preferences; projected lifespan of the digital material; and the widely varying carrying capacities (for capacity in 2001 see the Table on Commonly Used Types of Removable Media, below). However, for most projects, CD-ROM or DVD are the likeliest options, simply because users are more likely to have access to a CD-ROM drive or DVD drive. In addition, these media are fairly durable, and since they can be distributed in a read-only format, it is harder for users to modify or delete materials accidentally. In order to deliver data on removable media you need either the appropriate equipment for duplicating data onto blank media, or the means to outsource duplication to a vendor. You may also need an installation program (if the

applications do not run from CD-ROM). Finally, you need a means of delivering the media to end users and keeping track of orders and fulfillment.

Definition Box:

Commonly Used Types of Removable Media (2001)

Media Name	Full Name	Media Category	Storage type	Current Capacity
DAT/DDS	Digital Audio Tape/ Digital Data Storage	tape	magnetic	For DDS-3 12 GB uncompressed[1] and 24 GB compressed
DLT	Digital Linear Tape	tape	magnetic	35 GB - 70 GB
CD-ROM	Compact Disk-Read Only Memory	disk	optical	650 MB
DVD	Digital Versatile Disk	disk	optical	6 GB
Removable Disk Packs		disk	magnetic	10 GB - 70 GB
Floppy Disks		disk	magnetic	1.44 MB
Zip disks[2]		disk	magnetic	100MB and 250MB[3]
Jaz drives		disk	magnetic	1 GB and 2 GB

Networked delivery

Here the Internet or TCP/IP-based intranet are the chief options, although there are other types of network such as AppleTalk or IPX/SPX. There are a number of different delivery mechanisms to users, each suited to different purposes. HTTP[4] protocols are the standard mode for communicating web documents, and given the ubiquity of browsers they represent a near-universal method of disseminating text and images, without requiring special action on the part of the user. Streaming media —for example via RealAudio, RealVideo or MP3—allows you to supply real-time digital audio and video, without necessarily allowing the user to download and save the data locally. FTP[5], which supports file transfer only, is suitable for making raw data, including installable applications, available to users, who may download files and save them locally. In order to deliver data of any of these types via the network, you need the following: a network path between the provider of the data and the users; server hardware; server software (an operating system and server applications); and client applications enabling users to access the data (though as indicated above, this may simply mean free browser technology). If you prefer not to host the delivery yourself, you may be able to outsource it, either to an internet service provider (ISP)[6] or to a partner or consortium member with the appropriate infrastructure.

Pros and cons of portable media

The advantages of portable media for producers of digital assets center chiefly on their cost and reliability, and on the fact that they provide a mechanism to deliver very high quality content to users. The media (e.g. CD-ROMs) are typically fairly cheap. Sale of digital assets on physical media is relatively straightforward, presenting analogies with printed documents which may also ease copyright negotiations, and assets can be licensed multiple times, extending the possibilities of generating revenue. There are also few potential points of failure in the use of removable media, no need to rely on network connections, and no bandwidth restrictions. There are advantages to users too: data presented in one physical package is attractive, and users can be offered access to much higher-quality data than would be feasible over the Web.

There are a number of disadvantages of portable media for producers, many of them involving loss of control and information. With portable media, producers hand over their project data to the user in one convenient package; once released, the materials cannot be updated or augmented except by an additional publication. Furthermore, access control is difficult compared to networked resources, and producers cannot gather usage data or other information. Although they eliminate dependence on networks and hence can be valuable in environments where no network is available, portable media are vulnerable to loss, theft, damage, and physical deterioration. Their storage capacity—though increasing steadily—is nonetheless small compared to the size of a high-quality image archive, and is more suitable for delivering a focused sampling of materials than a large collection in its entirety. From the development standpoint, too, there are disadvantages. Largely because it is often heavily customized, application development for portable media has traditionally been more expensive than for Internet-based delivery, and multi-platform development can be much more difficult (although there are also challenges in supporting multiple browsers and access tools for Internet-based distribution as well). Publication of portable media also poses potential difficulties; for one thing, it requires the production, packaging, storage, and shipping of physical objects, which may be part of a museum's ordinary work but requires a kind of staffing and space which libraries or academic projects may lack. And as few museums have access to media distribution and sales channels, their products do not get the shelf-exposure that CD-ROMs and DVDs released by both traditional and new media publishers achieve. As a result, although a small number of cultural CDs have sold over 100,000 copies, most sell only a few hundred and are only distributed in the local museum store.

Pros and cons of networked delivery

The advantages of networked delivery for producers center on ease of development and production: if the project is well planned, a single source of data avoids the need for duplication of data (and this is a consideration to be attended to carefully); there is transparent updating and fixing of application bugs; many client and server components already exist ready-made; and multi-platform development is far easier than for removable media. It is easy to create tiered access tailored to different target audience needs, and measuring usage is relatively simple. In addition, browsers provide a useful

set of generic client capabilities. Institutions that subscribe to many digital resources—for instance, academic libraries—overwhelmingly prefer networked access over portable media because it simplifies their task immensely, eliminating the accession process, relieving them of the care of fragile physical objects, and broadening their users' access to the materials.

The disadvantages of networked delivery focus on the relative unreliability of networks and network applications: there are many potential points of failure; there are varying, non-standard, and buggy application implementations (Java, JavaScript, CSS, XML) and limitations in browser clients and interaction. Being on the Internet also raises security concerns, although these are being addressed in a number of ways. Charging for networked access is not as straightforward as the simple purchase of a CD-ROM, although again there are a number of workable models currently in use. Finally, the quality of the digital material that heritage institutions can make accessible via the Internet is constrained by the available bandwidth, and although bandwidth has been increasing steadily for the past two decades, so too has the demand.

For most purposes, some form of networked-based web application is a good solution, and this is the approach that most digital resource providers are now taking. Given the nature of digital collections, being able to update and add to the resource dynamically is a huge advantage for most content providers, as is the ability to add users effortlessly without producing and mailing additional disks and packaging. The Internet provides a direct distribution channel between the heritage institution and the user community. Most heritage institutions that use the Internet to make their digital assets accessible do so for free, perhaps because of current attitudes towards charging for networked information resources, although it is notable that licensed digital resources such as journals, full-text collections, and bibliographies are widespread and take an increasingly large share of library purchasing budgets.

Modes of access

Assuming a web-based, networked delivery scenario, the next set of choices concerns the application model for web delivery, and the underlying technologies and data structures that support the digital publication. As with other decisions, these will be strongly influenced by the scale and purpose of the project or program. The needs of a small-scale project with limited, unchanging content will be very different from those of a large organization with high-volume delivery needs.

The simplest form of web delivery is a static HTML page: essentially a document that has been encoded so as to be readable using a web browser. Pages of this sort reside on a web server and are delivered to the user for viewing when requested. They may be searched in very basic ways by various means: by using a search engine installed locally, or by using one of the available commercial search engines such as Excite or Google. Such pages may be quite simple to create and maintain, using standard HTML authoring tools such as DreamWeaver or Microsoft Front Page. Since HTML is a very basic

markup language, it is readily understandable and easy to learn and use. Furthermore, the software required to publish such pages is simple and easy to support.

Some of the limitations of this approach have already been indicated in Section V, Digitization and Encoding of Text; HTML has extremely limited descriptive capabilities and can represent very little of the intellectual structure or content of a document. Thus, while HTML is entirely adequate for the delivery of a project web site (home page, project description, contact information, background, and the like), it is not appropriate as a way of capturing the actual intellectual content of a site (primary source documents, metadata, bibliographic records, image collections, and so forth). But in addition, the structure of such a site—a simple collection of HTML files—does not offer any means of managing the data systematically, querying or displaying it powerfully, or maintaining it as it grows.

These two limitations—lack of descriptive power in the encoding, and lack of overall data management—can be addressed separately or together, depending on the project's needs and the nature of the data. The former can be addressed by adopting a more powerful descriptive markup, using an SGML or XML encoding system such as TEI, EAD, or one of the other high-quality encoding systems now available (see Section V for more details). There are now a number of free open-source tools, as well as commercial systems, for publishing XML documents on the web, using XSLT (Extensible Stylesheet Language Transformations). If the content to be published consists of textual items with rich internal structure that may be of research interest to users, XML is an effective way of representing this information and exploiting it in the user interface. Good examples of this kind of content are scholarly editions or full-text collections of historical or literary documents. Similarly, if the content consists of highly structured data such as bibliographic records, finding aids, or metadata, XML may also be a useful way of capturing and publishing it, albeit with a different kind of DTD. In all of these cases, you can use XML tools to create a publication which enables users to query the information in detail, view and manipulate the results, browse documents in their entirety, and perform other central research activities. At the moment, the requisite publication tools require considerably more technical competence than those for basic HTML; the XML markup is likely to be more complex, and the design and implementation of the XML delivery system require an XML programmer. However, there is a strong demand for XML publication tools, that allow a comparatively non-technical person to manage and publish XML documents effectively, and such tools may well start to appear within the next few years.

The second limitation, the need for more powerful data management systems, needs to be addressed at a deeper level: not in the encoding of the individual items, but in the creation of a powerful infrastructure in which they are stored and managed. Typically such systems are built on a database model, treating the individual items as database records and exploiting their regularity and structural similarity to allow for effective searching and manipulation. Such systems may be quite small—for instance, many web sites are based on small-scale database applications such as Access or even FileMaker, and for very limited purposes (such as publishing a small bibliography) these might well be an

improvement over static HTML pages. However, for large institutional projects, robust database systems such as Oracle, Informix, and MySQL (the latter is free and open-source) are more typical and are capable of managing very large-scale collections. What the database architecture provides is the ability to query large numbers of documents with maximum speed, and to perform the kinds of processing—sorting, statistical manipulation, systematic comparison—that database tools are best at. Such a solution is ideal for a large collection of similarly structured items, where retrieval speed is very important: for instance, large metadata collections, image libraries, historical census data, and the like.

Using a database architecture offers a number of advantages in addition to power and scale. Once the underlying design is complete, the database can grow almost without limit as new records are added, without any work other than that of entering the new data. The system can include an input form which allows non-technical staff to make updates or additions, and the form can be made accessible over the web to accommodate off-site staff or collaborators. The newly added information may become live immediately, or can be marked as provisional, to be published pending review by an editor. Such systems also offer powerful workflow management and tracking possibilities which can be a valuable investment for a long-term project or one involving several participants.

One of the most distinctive features of both the XML and database-driven approaches is that in both cases, the web page that the user actually sees is generated dynamically from the source data (the XML document or database records), rather than being stored statically on the server. In the case of the database, the publication system retrieves information based on a user query, and then uses a style sheet to order, format, and display the results in HTML so that it can be viewed with a web browser. Similarly, although XML documents themselves are stored in a stable form, the XSLT stylesheets which turn them into HTML for web viewing can also perform more powerful transformations, such as reordering data elements, adding text and graphics, selecting and displaying small segments of the document, or even converting the information to another data format entirely. These dynamic capabilities open up a huge variety of interface options which go far beyond what is possible using simple HTML.

Finally, as suggested above, these two approaches—XML encoding and database architecture—can also be combined to leverage the power of both, and indeed there are signs that the two technologies may be converging, as database systems become more complex and XML tools become more powerful.

In addition to the questions of technical implementation and user access discussed above, access models are also concerned with issues such as security, limitations on access and charging, which will be considered next.

Most of the projects interviewed in the survey for this *Guide* are currently providing free access to their collections, but many are looking at the alternatives or have already chosen a licensing model. Models for charging for the use of digital assets over the Internet are still not as widespread and straightforward as they might be (evidence of this

comes from the continuing battles over Napster-type applications). But considerations such as protection of rights, or the need to generate revenue from digital assets, provide increasing motivation for projects to limit access to their online resources. These limitations can take several forms which may be more or less burdensome to the user. Typically users need to identify themselves before gaining access, via an authentication system of some kind. Such systems may involve direct user action, such as entering usernames and passwords, but there also exist modes of authentication that are invisible to the user, such as IP address authentication or access via an authenticating gateway. Whatever method is chosen, restricting access can be costly, determined hackers can find loopholes in most systems (particularly if the restricted data is valuable enough), and authentication systems may require a high level of maintenance and technical support. Projects or programs need to be sure that the revenue generated will justify these costs and that proper account is taken in the project planning of the responsibilities and legal liability of the agency with control over the server that distributes the assets. The advantages of free access to digital materials are therefore not just altruistic — there could be a significant overhead associated with limiting access. Many of the projects and programs interviewed were looking at revenue generation as a means of sustainability, so it is likely that secure systems that are easier to implement will be generally available in the near future. Corbis and Getty Images both believe that future access to picture libraries will be via the Internet, and they are building libraries and systems to make this happen. Many cultural institutions have generated income from their existing collections through the licensing of analog copies, and they need to make the shift to digital delivery of the material.

Future Trend:

XML

The advantages and disadvantages of static and dynamic web pages have been outlined above. One possible distribution method that combines the simplicity of authoring and hosting of HTML with the scalability and structure of dynamic pages is XML as a data format in a native XML database or content management system. For further information see Ronald Bourret's sites:

http://www.rpbourret.com/xml/XMLDatabaseProds.htm

http://www.rpbourret.com/xml/XMLAndDatabases.htm

Structural Metadata

As the digital representation of our cultural and heritage material increases and becomes more complex the relationship of individual digital objects to each other, the item from which they were derived, the collection to which they belong and the way in which the digital objects are stored and organized will become increasingly important. It is this information that will enable future users to retrieve not just a faithful representation of an object but reconstruct, navigate and understand the whole context in which the object was created and used. For this to be achieved distribution systems will have to hold increasing amounts of structural metadata. This may also suggest holding such metadata in object orientated rather than flat file or relational database systems.

Metadata Harvesting

As digital objects, catalogues and finding aids proliferate on the World Wide Web, effectively searching and retrieving information across multiple servers, systems and domains becomes both increasingly important and increasingly difficult. This issue is a central concern for cultural heritage and humanities institutions, which share the goal of broadening access through digitization.

One solution to this challenge is Metadata Harvesting. Simply put, this is a protocol that exposes metadata on the World Wide Web to enable searching across repositories. The best known system is the Open Archives Initiative Metadata Harvesting Protocol (MHP).

Further details:

A non-technical introduction to MHP:

Clifford Lynch, "Metadata Harvesting and the Open Archives Initiative," ARL Bimonthly Report 217 (August 2001): http://www.arl.org/newsltr/217/mhp.html

and:

Donald Waters, "The Metadata Harvesting Initiative of the Mellon Foundation," ARL Bimonthly Report 217 (August 2001): http://www.arl.org/newsltr/217/waters.htm

The OAI MHP protocol: http://www.openarchives.org/OAI_protocol/openarchivesprotocol.html

MHP tutorial: http://library.cern.ch/HEPLW/4/papers/3/

CIMI Working Group: http://www.cimi.org/wg/metadata/

CLIR Metadata harvesting project: http://www.clir.org/activities/details/metadata-docs.html

DLF and Metadata Harvesting: http://www.diglib.org/architectures/mdharvest.htm

University of Illinois at Urbana-Champaign Metadata Harvesting services: http://oai.grainger.uiuc.edu/

Conclusion

In determining the best delivery options, user needs should be balanced against producer advantages. Though portable media are cheap, reliable, relatively easy to sell, with no bandwidth problems and appear easy to archive, network delivery is generally easier and cheaper to produce and update. Tiered access for different audiences and for measuring usage is simple online. Despite bandwidth restrictions and security concerns, for most purposes networked delivery is the most common solution, providing a direct distribution channel that often reaches unexpectedly wide audiences.

Depending on the scale and purpose of a project or program, a database-driven system will generally be preferable to static HTML pages. Although new developments, such as metadata harvesting, are making content stored in databases more easily discovered, HTML pages are still more easily searchable from outside than dynamically generated pages, although within the resource itself the reverse is true. XML pages, though more costly to develop, are more economical and easier to maintain and update, and they offer vastly improved possibilities for data description, manipulation, and reuse. As with other decisions, this one will also depend on how users use your material as well as the scope of your project.

Access models also have security and economic aspects. Rights management and the need to generate income from digital assets may mandate access limits via identification and authentication systems, which, however, add another layer of cost for development and maintenance. New security and economic models are important to watch for.

---

[1] Because of the way data are encoded when written to tape and the way error correction works we do not recommend that information be written to tape only in uncompressed format.

[2] Zip and JAZ are proprietary systems manufactured by Iomega.

[3] Note the 250MB drives cannot read 100MB media and vice versa.

[4] Hypertext Transfer Protocol: A set of rules that enable documents to be used on the worldwide web and viewed by browsers such as Internet Explorer or Netscape.

[5] File Transfer Protocol: enables files to be transferred from computer to computer using a simple program such as WSFTP or Fetch.

[6] Internet service providers most typically provide internet access for users. But many of the larger ISPs also offer server services which can accommodate the publication of large web-based resources, for a fee. If your institution is not able to maintain its own web server, or if the resource you are creating is too large to be published from your local servers, an ISP may be an option worth considering.

XI. Sustainability: Models for Long-Term Funding

Introduction

Regardless of the quality and robustness of the digital resources created, they will not last long, let alone grow and survive in the long term, if the projects that produce them cannot be sustained. Sustainability thus refers to all of the considerations that go into maintaining the institutional context for the creation and maintenance of digital objects and resources, and supporting their long-term viability. Preservation—the maintenance of the digital objects themselves—is a distinctly different but related topic, dealt with in Section XIV. Digitization of cultural objects is still relatively new; the majority of projects and programs interviewed are less than ten years old. We are not yet able to take a long-term view of the life cycle of a digitization project, hardly even a medium-term view. In most cases, energies are naturally channeled towards start-up and building a critical mass. Although there was some recognition among projects and programs that a strategy for sustainability was desirable, it is probably true to say that sustainability in the long term is still generally pushed down the list of priorities by more immediate and pressing concerns. Readers should also consult Section II on project planning, which puts these issues into the context of the larger planning process.

Few projects have identified a point when, with their work complete, they will close down and pass the content they have created in digital form over to others for long-term maintenance and delivery. And yet for many projects, whether they plan for it or not, this is the likely outcome, either because of loss of funding or simply because the project has finished its work and has no further reason to continue. It is worth considering at the outset whether the digitization work you are undertaking has a finite scope, or is likely to continue indefinitely. Projects that are defined by the scope of a particular collection or author should plan for an appropriate exit strategy when their work is complete, including plans for ongoing delivery of the digital resource created.

The Survivors of the Shoah Visual History Foundation, for instance, advises that project managers plan out the whole system before any digitization takes place, envision the entire process and life-cycle of the project, and plan for it up to and including preservation of the materials. This plan should ideally include contingencies for re-locating the digital objects if, for example, funding declines to a point where it is no longer viable to maintain the materials.

Projects whose mandate is broader will be limited (or not) by other factors, such as institutional priorities and funding. For some institutions, such as national libraries, it is entirely appropriate that they should not have an exit strategy: the mission and funding of such institutions is consistent with long-term support and maintenance of the digital materials they create. But it is not only the national libraries that have such far-ranging plans: some large academic institutions are also making a commitment to keep digital

materials over the very long term. The Robertson Media Center at the Clemons Library, and the Special Collections of the Alderman Memorial Library, both at the University of Virginia, intend to keep their digital deliverables viable and available indefinitely. The University of Virginia Library as a whole is launching a Central Digital Repository that will, in theory, become the sustainable home for projects currently delivered via one-off systems at the Robertson Media Center and Special Collections. The Harvard University Library Digital Initiative (LDI) adds another aspect to the strategies mentioned above. Repository policy strongly encourages deposit of at least one version of the digital object in an approved preservation format to ensure that new deliverables can be generated affordably in the future.

Most projects reported that loss of the digital materials would matter greatly—institutionally and to their user communities—but few could be confident of the continued stream of funding and other support that would be necessary to maintain their digital assets. There are high costs associated with ensuring long-term access, for example: regularly updating the interface, maintaining backup copies, migrating the data to new platforms and formats, and supporting the use of the resource, let alone providing for its continued incremental development over time.

Programs and projects need to think more strategically, and need to adopt a longer-term view by planning out the development of sustainability plans at the outset of their initiative. Funding agencies have a role to play in insisting that provision for long-term sustainability be factored into grants They also have a wider responsibility in recognizing the cultural value of digital objects and contributing towards their sustainability. However, long-term sustainability is difficult to achieve on grant funding alone, and a reliance on “soft” money puts a project on a very insecure footing: not only because of the uncertainty of funding from year to year, but also because so much of the project’s resources must go towards the ongoing search for support. If funders value the digital cultural objects they can help maintain their longevity by encouraging the establishment of longer-term, strategic funding plans. Needless to say, the prevailing economic and political climate is relevant and not necessarily conducive to a strategic, long-term view.

In this section we will look first at the funding issues in more detail, and then discuss planning strategies and methods of sustaining resources through use.

Funding

Central Funding

The Library of Congress’s National Digital Library Program (NDLP) intends to keep its digital deliverables available for as long as possible. The adoption of an exit strategy, in case the Library found itself unable to meet the costs of sustaining the NDLP’s digital assets was considered unthinkable. On the contrary, the Library recognizes that it is more likely to provide an exit strategy for other institutions that find themselves unable to sustain the digital resources they have created.

For the Library of Congress, the sustainability of its digital resources would not depend upon their generating income streams to make them self-sufficient. In our interviews, the Library reported that it had already obtained some funding towards providing a level of sustainability for its digital assets and is working to make these resources visible to national audiences, thereby building a large group of invested stakeholders.

Similarly, the National Library of Norway's overall statement of intent, or mission statement, refers to the need to establish a framework for digital libraries and collections that make information available to everybody, wherever they are, in a democratic way. The emphasis is on access and inclusion. They intend to establish a long-term digital library with core functionality, and then provide different "windows" onto this information. As with the Library of Congress, there is no exit strategy — on the contrary, the digital activities at the Library aim at building a sustainable digital library, with sustained and secure funding.

Museums face the same challenges in finding opportunities to fund digital collections project–by project and collection–by collection. While the Berkeley Art Museum/Pacific Film Archives builds its digital projects with sustainability and preservation in mind, it must fund the creation of new digital collections in exactly this way. Early digitization efforts at the Peabody Museum of Archaeology and Ethnology at Harvard University also followed this methodology.

Base vs. project funding

Many projects, like the Cornell Institute for Digital Collections (CIDC) at Cornell University have benefited from very generous lead-gift funding and continue to rely on soft money. However, the use of grants and gifts to fund digitization initiatives means that the prioritization and selection of material for digitization may be driven by an assessment of those areas most likely to receive funding at any given time (through their timeliness or appeal to a particular funder), rather than by an overall digitization strategy. These projects recognize the need for a stable funding environment.

Similarly, the Virginia Center for Digital History (VCDH) at the Alderman Memorial Library at the University of Virginia (UVA) has benefited from funding from a wide variety of sources including UVA, National Endowment for the Humanities (NEH) and private donors. The main problems with the funding have been uneven distribution and inconsistency. With the benefit of hindsight, the Center felt it would have been better placed if it had had solid base funding with project funding added on top.

Securing longer-term funding

There are three primary ways to secure long-term funding for digital assets: commitment by the originating institution to meeting the costs associated with the digital asset, endowment funding, and making the resource self-sustaining. Of these, endowment funding is the most difficult to obtain and none of the projects interviewed for this *Guide* described a successful approach of this sort.

Several projects, however, have set in place a funding model that relies on generating ongoing support from the resource itself. The Brown University Women Writers project is dependent upon self-generating funds; it has looked for endowment funding but now hopes that the resource itself will generate sufficient income to meet its long-term costs. The sustainability of the Colorado Digitization Project also depends on self-generating funds and it is considering establishing a membership-based organization to generate revenue in a coherent and consistent way. In making its resource self-sustaining, JSTOR has established a basic fee structure for participating institutions: a one-time Archive Capital Fee (underwriting digitization costs and an archives reserve fund) and an Annual Access Fee (supporting ongoing access to the database and the addition of new volumes to the database). JSTOR director, Kevin Guthrie, comments that nonprofits should develop revenue sources that "match the nature of the mission-based uses of their funds." As JSTOR's work of archiving journals continues, so stable and recurring funding is essential. JSTOR is thus building a reserve fund, the proceeds from which will help fund the ongoing cost of archiving.[1]

Some plans for self-maintenance face greater challenges. The Thesaurus Musicarum Latinarum (TML) project is dependent for long-term sustainability on self-generating funds, but the TML has not yet secured these resources. Should resources prove insufficient, the exit strategy would be to break up the TML and the larger consortium group, and move them elsewhere. Another possibility would be to publish the TML from a private server. The TML experience suggests that projects need to recognize that even institutional support is tenuous, that digital data are easily destroyed, and that some projects will not survive.

Cooperation and joint initiatives with other types of institutions or agencies may also provide a route towards longer-term funding. Where several institutions combine to support the creation of a resource, the burden for each may be quite manageable. Similarly, funders may look more favorably on joint efforts that show a broad institutional commitment to the resource. The CIDC at Cornell, for example, has cooperated with libraries, museums, and archives, at all levels from local to international. While these are useful processes, cooperation is not without problems and firm rules have to be agreed to produce a successful project. However, cooperative agreements of this sort can provide a route to funds that are otherwise inaccessible. A more detailed discussion of inter-project collaboration and its benefits and risks can be found in Section IX.

Effective strategies and planning towards sustainability

The Digital Imaging and Media Technology Initiative (DIMTI) at the University of Illinois is an excellent example of "joined up thinking" in relation to digitization. The benefits of establishing a body such as DIMTI at the outset of digitization are demonstrated in the strategic overview that it enjoys. The fact that it is a single body with responsibility for dealing with generic issues from the outset has advantages in creating a unified institutional vision. Furthermore, DIMTI demonstrates that such a body does not need to be large when supported by other appropriate administrative or managerial bodies. The fact that the University of Illinois Library has a relatively flat structure seems to have been beneficial in this case.

Although various forms of base and soft money remain the predominant forms of sustainability, the projects surveyed for this *Guide* tended to be large and relatively well established. Therefore, they may be generally less concerned about their ability to secure future funding than smaller and more recent projects. However, some projects have developed alternative models for sustainability.

The licensing model developed by the JSTOR project secures long-term funding by creating durable relationships with other stake-holders: most significantly, the collection's current and future users. In this model, access to JSTOR collections for UK higher education institutions is licensed via the UK Higher Education Funding Council's Joint Information Systems Committee (JISC). The JISC succinctly summarizes the benefits to both parties as follows: "the agreement between JSTOR and the UK Funding Councils (acting through the JISC) is a twenty-five year Agreement commencing on 1st August 2001 and terminating on 31st July 2026. The length of this agreement is to reflect the on-going partnership between JSTOR and the JISC, to allow institutions to feel confident in long term access to JSTOR collections, and to underscore the value this resource brings to the UK academic community." Such sub-licensing agreements are effective methods for projects that have a large and identifiable institutional audience to secure funding and sustain the project through long-term use.

The Colorado Digitization Project's collaborative, not-for-profit model is a second alternative. The project has a broad range of funders (nine in total) and participation is open to libraries, library systems, archival organizations, selected pilot schools, historical societies, and museums. The involvement of more than 50 collaborative projects and four regional scanning centers, and the emphasis on sharing and transferring expertise means that sustainability does not depend on a single funder or institution but is shared by the collaborative partners.

Neither of these strategies is guaranteed to ensure sustainability. Licenses may not be renewed or obtained in sufficient numbers and contracts for large consortia, particularly regarding ownership of material, are difficult to achieve. Nevertheless, the demand for grant-based funding will always outstrip its supply and such models indicate alternatives that can go some way to achieving sustainability.

Focusing resources and explicit statements of intent

Focusing resources on a clearly identified target audience has been shown to be helpful in sustaining projects, as that audience has an important stake in the future sustainability of the program. The UVA Electronic Text Center reported that the most effective dissemination strategy for building a sustainable resource was predicated on the Center's local focus on research, teaching and learning, and the grounding effect this produces. Similarly, the William Blake Archive (also at UVA) has produced an explicit statement of intent that covers its rationale, scope, significance, primary audience, long-term sustainability, level of faithfulness to the originals and suitability for different target audiences. This level of focus has enabled it to sustain a high degree of momentum and hence of attractiveness to its audience, which in turn strengthens its funding position.

Sustaining resources through use

Ensuring internal and external use of a resource can contribute to its sustainability. The Walker Art Center (WAC) and Minneapolis Institute of Art (MIA) both intend to ensure that the digital process becomes part of their institutional culture, ensuring that all types of user use the resources from day to day. They will educate staff to identify and use the objects in everyday work, and if they are successful, this integration of digital materials into the mainstream of users' processes will help ensure the sustainability of the digital objects. Similarly, DIMTI is an example of a program that has set itself the goal of integrating digitization into existing Library activities and structures; this is a far-sighted and effective mechanism to garner institutional support and maintain digitization activities in the long term. The sustainability of digital resources will be increasingly assured if academics are successfully encouraged to use them.

Summary Box:

Key elements in establishing sustainability:

- Effective strategies and planning towards sustainability — ensuring that information about your resources is easily available and that potential users are encouraged to adopt the resources (see Section II, Project Planning);
- Use of appropriate standards to facilitate the sustainability of resources by making them less susceptible to changes in technology, and easier to migrate (see Section XIV, Preservation);
- Developing groups that are stakeholders in your resources at the outset (see Section IX, Working with Others);
- Getting into the mainstream — by ensuring the widespread use of digital resources in learning and teaching as well as research (see Section II, Project Planning);
- Migration of skills as well as technology.

Migration of skills as well as technology

The Genealogical Society of Utah, Family History Department reported that it is not only technology that ensures the longevity of materials, but also migration of hardware, software, media and, importantly, people skills. The GSU is constantly reviewing its management procedures to ensure that no information can become obsolete due to hardware, software, media or skills being lost. (See the discussion of skills in Section II, Project Planning.) Nearly all projects will eventually need to migrate their digital resources to new environments and will require funding to do so, but these costs are rarely built into the project's base funding.

Building and sustaining digital collections: NINCH/CLIR models

Cultural heritage institutions face a number of challenges when placing digital collections online. Not least of these is the problem of how collections can be built and sustained in harmony with the sector's non-profit culture. In February 2001 the Council on Library and Information Resources (CLIR) and the National Initiative for a Networked Cultural Heritage (NINCH), supported by the Institute of Museum and Library Services, convened a meeting to discuss what sustainability models were available to the sector. The meeting included representatives from museums and libraries, legal experts, technologists, funders and web enterprises.

In discussing business models the group concluded that whether the digital collection is for profit or not, large amounts of capital, new skills and new organizational cultures are required. Furthermore, non-profit organizations must adopt the same rigor for accountability, efficiency and value for money as commercial enterprises. The latter may have a competitive advantage in their ability to identify, test and target audiences, but any successful project must have clearly identified its purpose, mission and audience. Nevertheless, non-profit organizations generally have a longer period of time to realize their goals, and the expectations of trustees and governors can differ considerably from that of shareholders or investors. For the report, *Building and Sustaining Digital Collections: Models for Libraries and Museums*, see http://www.clir.org/pubs/reports/pub100/pub100.pdf

It is becoming apparent that web access to digital collections is not just broadening access but providing access in new ways. This presents a challenging dichotomy for non-profit organizations. How can the convenient and unrestricted access to digital collections that users expect be reconciled with the need to recover costs? Moreover, the resources, outlook and skills required to build a digital collection are somewhat different from those required to distribute it. The non-profit organization may find that the latter are far harder to integrate into their larger mission.

Cultural heritage organizations could license their digital content to a commercial distributor, but it is by no means clear that the analog/digital divide is the point where non-profit organizations cease to owe users free access. Even though access to analog

collections has never been entirely free, it is still uncertain at what point—and to what extent—cultural organizations should start using web delivery to load costs onto the user.

An alternative model is strategic collaboration between non-profit organizations. (One such model examined at the CLIR/NINCH meeting was between the George Eastman House, Rochester, New York, and the International Center of Photography in Manhattan. See page 7 of the report cited above). The costs are not necessarily going to be lessened through this approach but it does provide a means of spreading and managing the risk involved, building economies of scale and meeting the need for standards and practices for interoperability and access. However, as noted above, even between large organizations such collaboration is no guarantee of a successful return. Considering the infrastructure and skills required for creating and delivering digital content, small and medium sized libraries and museums are likely to find that collaboration with a larger partner is the only method available to them.

Despite the risks and costs involved, participants at the meeting voiced the opinion that not to enter the digital arena at all was itself a greater risk. At least the risk could be shared by working collaboratively. Furthermore, the burden of expectation might be diminished if funders viewed as experimental some ventures where the return might be non-existent or elusive.

The elements of sustainability identified at the meeting require further work by the sector. Participants emphasized the need for standards and best practices, coherent and common digital architectures, and a means of creating and sharing knowledge. In order to deliver these desiderata, the group suggested a number of ways forward, including guides to best practice, a clearinghouse for technical information, a registry of digital conversion projects, digital service bureaus, regional centers for digital content production, and "tool kits" for data and communication design and development.

---

[1] "Developing a Digital Preservation Strategy for JSTOR, an interview with Kevin Guthrie," RLG DigiNews (August 15, 2000) http://www.rlg.org/preserv/diginews/diginews4-4.html#feature1. See also the section "Who pays what," in Dale Flecker, "Preserving E-Journals," D-Lib Magazine (September 2001) http://www.dlib.org/dlib/september01/flecker/09flecker.html

XII. Assessment of Projects by User Evaluation

Introduction

There are two main reasons for carrying out evaluations: to ensure that digitization initiatives produce materials that meet user needs, and to assess whether the original objectives of the project were achieved. Throughout this *Guide* we have stressed the substantial and long-term financial investment that digitization requires, and have drawn attention to the demands it places on limited organizational resources. It makes good sense, wherever possible, to assess the results methodically and make the benefits of this investment evident.

Evaluation provides a mechanism to measure objectively what has been achieved through digitization. Funding bodies increasingly require that digitization projects deliver an evaluation and impact assessment report as part of their final report. In addition to helping projects to be more accountable to the user community, evaluation provides an opportunity for projects to discover who is using the digital collections and how. We make many assumptions about our users and their needs, often basing our expectations of digital usage patterns on existing usage of analog resources. However, at best, these can only be educated guesses, since access and usage can change completely in the networked environment. For instance, use of postcard collections has always been limited, but when made available in digital form their use rises dramatically; ease of access is the key. Only by carrying out evaluation with our users can we find out how the digital resources we create are actually being used. Despite the spreading use of digitization in the cultural sector, we know surprisingly little about its effect on the various user groups we aim to serve and the way digital collections influence their teaching, research, life-long learning, personal development, or entertainment. Research and evaluation will help indicate how we can enhance the functionality and usability of the digital resources we are creating.

At a more practical level, giving the target user group the chance to evaluate and test the delivery system, and in particular the interface, at an early prototype stage should enable the development program to take advantage of feedback and suggestions for improvements. This approach makes most sense when projects and programs have built in the resources needed for iterative prototyping, so it is a good idea to ensure that you have staff and time to implement revisions if you get early feedback from potential users. Commercial and cultural institutions report that this approach allows them to be more responsive to a broader spectrum of users with a variety of technical experience and familiarity with the subject. User evaluation and research can provide answers to these (and other) questions:

- Who uses the digital collections? What are their demographic characteristics, interests, level of computer skills?

- How are the digital resources being used? What navigation paths are followed? What are the popular choices, the types of searches performed?

- How does the use of the digital collections relate to that of the analog ones? Are they complementary? Is the profile of the users for each category different? Has usage of the analog resources changed as a result of digitization?

- What is the effect of the digital resources on the users? Are they satisfied with the service and collection? Does the resource trigger interest in new topics, or challenge existing perceptions? What are the formal and informal learning outcomes?

The questions you ask will depend on the aims of the program and the particular focus of your interests. For example, if you are interested in educational uses of the digital surrogates of a museum's collections, the main issues to explore might be the kinds of understanding and learning the digital resources can enable, and how you can ensure that visitors who use them also engage with the original objects on display. If the digital resources are to be made accessible in a museum gallery, then evaluation activity might examine how computer interactives affect the visitors' experience, and how they affect the social context of group visits and interaction between participants in the group. In an academic library setting, evaluation might assess whether digitizing a given collection is likely to address local curricular needs, and what kinds of improved pedagogical outcomes might result. Evaluation can also help discover how effectively digital surrogates represent the original. Do visitors/users prefer to see the actual object or do the digital surrogates have the same impact? Questioning the user base can reveal this and any evaluation activity should approach this issue.

In making these assessments, remember also that many resources will be used by very different user groups: users from a variety of age groups, educational levels, backgrounds, and interests. Their needs and expectations will also differ considerably. An essential starting point before beginning any evaluation is thus to categorize the user community to make the assessment more informative. For example it might be useful to group users by:

- their subject area of interest (e.g., Classical Greek archaeology, Abstract Expressionist paintings, the history of Ethiopian Jews);

- their age group and educational level (K-12, college student, faculty, life-long learner);

- their function or role (e.g., librarians, conservators, curators, students, general public);

- the use they would make of the resource (e.g., research, teaching, life-long learning, casual browsing).

You can further refine this categorization by considering the potential user communities with reference to the project's defined context and priorities. For instance, it may be important to ascertain whether specialists actually use all the extra search facilities they say they want, or whether they use digitized resources more or less as the general public do. You may discover that different communities actually use the resources in very similar ways. On the other hand, certain users might have particular requirements, for example in relation to image quality or searching facilities.

Types of evaluation

Evaluation is classified into three types according to the point during the lifecycle of the project at which it is carried out: front-end, formative, and summative.

Front-end analysis is carried out before a program or application is developed. This type of evaluation can gauge potential users' reactions to the subject matter; assist in the selection of content areas and themes; provide feedback about the type of service and functionality that would be suitable; and enable the project to make a general assessment of the attitudes that stakeholders and users have towards the proposed development. Involving both groups early on helps ensure that the final product or service conforms more closely to their expectations and needs. *User needs assessment* often plays an important part in front-end analysis. This technique investigates the requirements of target users—(search tools, image or sound quality, depth of metadata—using methods such as focus group discussions or interviews, described below). The front-end evaluation process should also examine whether digitization is the best way to achieve the aims of the project or whether simpler and more affordable solutions might be more appropriate.

Formative evaluation takes place during the development phase and its results help refine and improve the design and delivery of the resource. It can be used to test the appropriateness and intuitiveness of the user interface and pinpoint problematic areas and programming errors. This is a vital step in the design and development of all digital resources. Even if the final product is not perfect, it will be better than if no user testing were carried out at all. It is never too early to test and involve future users in the design process. Even mockup screen views drawn on paper, sketchy web pages, or crude prototypes can provide valuable feedback and suggest changes before too much time and effort have been expended. Where a large user sample is too difficult to administer, even a brief survey with a small sample number of users will, in most cases, offer useful information at this stage. Formative evaluation also provides an opportunity to assess users' perceptions of the project's content. Different user groups might provide crucial feedback about the kinds and quantity of contextual information, metadata and tools for using the collection that they would find useful. The project will quite possibly discover at this stage that different groups of users have different expectations and may find that there is a need to narrow the focus of the project rather than attempt to provide the definitive resource to all potential users. Since in many cases digitization is a continuing

effort for cultural institutions, formative evaluation should also be an ongoing activity, integrated into the digitization chain and implemented as part of the working processes.

Summative evaluation measures the effect and impact of the completed program or a distinctive stage of its development. It is good practice to define particular stages at which to conduct summative evaluation—for example, after a particular phase of activity or once a particular deliverable has been completed—especially where digitization is an ongoing activity. Summative evaluation is often more thorough and yields more determinate results, involving real users as opposed to targeted potential ones. Since it takes place after all the selected materials have been digitized and associated interface and tools designed, it offers a more accurate picture of how these are perceived and used than that provided by formative evaluation. When the whole range of resources is made available, interesting observations can be made, for example, about the relationships between them, the most and least popular materials, and the associations users make between different materials. It may also be possible to find out why users find some materials of greater interest than others.

Often the summative evaluation is the first time that evaluators can measure in depth the effectiveness of interpretative exhibits and gallery kiosks, in relation to the surrounding space, and in the context of the exhibition itself. This is also the opportunity to explore the dynamics between real and surrogate objects, visitors, and computer interactives. The approach and tools chosen for summative evaluation will naturally depend on the aims of the survey and the reasons for carrying it out.

Evaluation methods and tools

Measuring and recording the effect of digital resources in cultural and educational settings, as well as the personal meanings that people derive from them, is a complex and difficult undertaking. Gay and Rieger (1999) have argued convincingly that "[m]edia- and technology-rich environments, such as cultural Web sites, demand equally rich data collection and analysis tools that are capable of examining human-computer interactions." There is no single golden method for evaluating digital programs and measuring their effectiveness and impact. Experience has shown that it is better to combine several methods (known as 'triangulation') in order to verify and combine data, relating quantitative with qualitative results. The dichotomy between the quantitative and qualitative approaches is usually artificial, since they work best in complementary ways, illuminating different aspects of a complex phenomenon.

Several useful pointers have emerged from experimental work in evaluation:

- Feedback from real users of digital collections and programs in their 'natural environment' (ethnographic research or the naturalistic approach) is crucial and usually more useful than testing in controlled situations with users solving tasks which evaluators have pre-defined.

- It is important to examine the use of the technology in its social context, rather than relying only on technological measures.

- As with user surveys in general, it is essential to define and follow a rigorous sampling methodology in order to acquire valid results that allow inferences to be made about the whole population of users. (Random sampling ensures that every person in the user population has an equal chance of being selected to complete a questionnaire or offer information, but this may not be the best method. For example, if you wished to gauge the effectiveness of the materials used by elementary school children, a stratified random sample might be more appropriate.) When collecting your sample, try to spread the process out over different days of the week and times of the day.

- Don't forget the non-users! Why do potential users not use your digital collections? Is it because they lack access to adequate technology or because the deliverable and the way it was presented did not appeal to them? What are the biases and barriers against the use of your digital collections? The answers of non-users can give you useful food for thought.

Some of the most commonly used methods in traditional evaluation work can be applied to digital programs, and these are briefly described below.

Computer logging of user interaction

Automated logging of user interaction with a digital resource provides a reliable way of recording users' choices and the path they selected through the web site or program. There are numerous programs for recording web usage statistics, several of which are freeware or shareware.[1] Most of these offer possibilities for graphic displays, diagrams and categorization of results according to various parameters (e.g., requests by day, month, year). They usually record the number of requests made, the folders and files requested, and a list of countries or types of sectors from which the users come, often based on an analysis of the IP address of their computer. By measuring the length of time between selected links, researchers can estimate approximately the amount of time spent on individual pages. Although web statistics are notoriously unreliable and difficult to interpret (for example, estimating the number of individual users from the number of requests), they can still be very useful in a comparative way, even if the absolute figures should be treated with caution. In general, web statistics programs offer insufficient information to build a profile of the users beyond an estimation of where their computer is based. More detailed and customized information can be recorded, using specialized tools developed in JavaScript, Java, Visual Basic, or C. Research into scripting and computer logging is being carried out at Virginia Tech (http://www.cs.vt.edu/research).

Once the scripting has been set up, computer interaction logging is generally an easy and objective way of obtaining a large set of data, which can be analyzed statistically. One problem is that when this method is used for programs in public access areas, it is sometimes difficult to differentiate between the interactions of different users. Another is

that although it reveals the most popular user choices, it does not explain why they were chosen. The results are not very meaningful on their own, but can be useful when combined with interviews, focus group discussions and observation.

Sites that require user registration offer a richer range of possibilities for evaluation and contact with the users, although here evaluators need to take into account such issues as privacy and data protection (see Evaluation and Privacy Issues below). Registration requires users to register and then log in to the site using the same user name and password on each occasion. This provides more accurate information and can help develop trend data: for instance, to what degree use by a particular category of patrons may be rising or declining over time, whether particular types of material tend to be accessed by specific user groups, or whether certain user groups use particular types of material or explore the resource in particular ways. Although registration may deter some users, it helps if the registration page is carefully designed, with clear information about the purpose of registration and assurances as to how the data collected will be handled.

Techniques similar to web usage logging — and often more accurate — can also be used with stand-alone multimedia programs or CD-ROMs.

Electronic questionnaires

Although the results generally pose problems for valid statistical analysis, as the sample is self-selected, electronic questionnaires provide an easy way to obtain feedback from end users. They work by encouraging users to answer questions about the resource and their use of it, often by typing or clicking on multiple choice answers. Common practice is to require the user to make an active choice to complete the questionnaire, but more recently institutions, eager to better understand their users, have implemented short evaluation questionnaires that appear automatically a few minutes after the user has entered the site. Although more intrusive, this usually generates a high number of responses. The results of the questionnaires can be automatically loaded into server-side databases for subsequent analysis.

Questionnaires can also be sent by email or mailed to users whose email and postal addresses are known or who have provided this information. This allows for more flexibility and customization compared with the standard electronic questionnaire approach. Attractive and clearly laid out printed questionnaires placed next to the computer terminals can encourage local users to leave their impressions and comments about the program. Providing enough pens and visible boxes or assigned locations for returning the questionnaire can help increase the number of responses. Again, this method is not statistically valid, as it records answers from a self-selected and not necessarily representative sample. It might be a useful option for recording information about local users, as long as it is not the only method of evaluation.

Observation and tracking

Observing how people use digital collections in the classroom, gallery exhibition, reading room or public space can be very illuminating. It provides an opportunity to collect information about the physical, social, and intellectual contexts that affect the use of the digital resources, indicating, for example, the relationship with real objects in the gallery, the interaction between groups of users, or the ways students make connections between primary sources and the school's curriculum. Observations can be recorded on data collection sheets, with circulation paths or with checklists of specific behavior categories, together with personal notes. Video recording and web cameras offer alternatives and can produce a wealth of data, although these often take longer to analyze, thus raising the cost. Observation raises privacy issues, which we examine below.

Interviewing and focus group discussions

Interviews and discussions with a small number of targeted users provide an effective method of evaluation and offer the opportunity for both structured and open-ended data collection. Discussions with focus groups are often held during front-end analysis, as well as at other stages of evaluation. These are participatory sessions with small groups of people, from the real audience or a particular targeted subset, who are encouraged to express their opinions about the digitized resources and how they are using or would like to use them. Interviews and focus group discussions can be open-ended, with the interviewer or discussion moderator interacting freely with users, or follow a pre-defined set of questions. Open-ended interviewing can be particularly useful for front-end analysis, to test how and what the targeted audience thinks about a topic before beginning program development. If an application is intended for specific groups (e.g., a researcher's resource or a schoolchildren's outreach program), discussions with focus groups can be very useful during the planning and development stages. These preliminary discussions often help outline a list of questions for more formal interviewing. Interviews usually provide many useful and meaningful data but are time-consuming to administer, demanding on the interviewer, and difficult to analyze and categorize. Projects should consider using ethnographic software for annotating transcripts of interviews with users. This software permits the researcher to aggregate comments with a given annotation label.

When testing a prototype, interviewing and observation can take two forms (often described as 'cued' and 'uncued' testing). Cued testing involves explaining to users what the program is about and asking them to perform specific tasks or to answer questions. Another possibility is engaging users in conversation and encouraging them to 'think aloud' as they go through the program, while recording their responses. With uncued testing, users are observed unobtrusively as they use the program and are then asked questions about their experience.

Checklist Box:

Checklist of criteria

Identifying appropriate criteria is vital for every evaluation. These depend on the scope and purpose of the study, the aims of the digitization program, and the time and funds available. This checklist outlines some of the basic aspects of digital resources that can be assessed. Each project must develop criteria that provide appropriate measures reflecting the goals of their own program and its users.

User interface/delivery

- Is the user interface consistent and appropriate to present the subject matter to the users?
- If icons are used for navigation buttons or commands, do the users understand them?
- Is the quality of graphics, images, sound, video adequate?
- Is the text legible (fonts, sizes, layout, spacing)?
- Are the media successfully integrated?
- Is the level of interactivity appropriate for the intended audience and environment?
- Are all the interface elements presented appropriately on different computer platforms?
- Can the program be used effectively by disabled users? Does it conform to ADA requirements for disabilities?
- Does the delivery of the program cater to different learning styles and types of audience?

Structure/navigation

- Is the structure of the various components appropriate to the content? (linear, hierarchical, network, combination)
- Is the resource easy to navigate? Does it indicate the user's position, prior moves, and available paths?
- Does the type or depth of indexing match the content adequately?
- Is the method of labeling content adequate?

Programming

- Are there any programming problems or errors (e.g. dead-ends and broken links)?
- What happens if the users do not use the application in the way it was intended?
- Does the application allow for user mistakes?
- Is there feedback given during processing operations that may take a long time?

Content

- Is the number of digital objects adequate?
- Are the amount and depth of accompanying information about the objects adequate?
- Is the depth of indexing well matched to the content?
- Is the information accurate?
- Is the information complete or are there important parts missing?
- Is the information about the whole project, program, service or the individual collections appropriate, correct, and clear?

Overall impressions

- Does the resource provide useful information, arouse curiosity and interest?
- Is it easy to use?
- Does it attract users? What type and how many are using it?
- Does it hold attention? How long are they using it for?
- Does it fulfill its intended purpose? (e.g. does a research resource assist researchers effectively? Does an interpretative museum application help users get to know and better understand the collections?)

Link Box:

There are a number of resources providing guidance in the planning and design of evaluation strategies. Some of these are listed here:

- The Institute of Museum and Library Services (IMLS) in Washington, DC has a site on 'Outcome Based Evaluation', which offers an overview of this type of evaluation with other resources and links: http://www.imls.gov/grants/current/crnt_obe.htm.
- Professor Bill Trochim's Center for Social Research Methods (http://trochim.human.cornell.edu/) at Cornell University and the American Evaluation Association (http://www.eval.org/) have web sites that offer useful information on evaluation and social science research methods.
- Jakob Nielsen's site on Heuristic Evaluation (http://www.useit.com/papers/heuristic/) offers information about the evaluation of a user interface design.

Evaluation and privacy issues

Projects should be aware of the IRB (Institutional Review Board) protocol, designed to safeguard rights and welfare of human research subjects. The main points to consider are:

- Risk to subjects
- Selection of subjects
- Informed consent
- Safety of subjects
- Privacy and confidentiality

Full guides can be found at http://ohsr.od.nih.gov/info/einfo5.php3.

Some of the evaluation methods proposed may be seen as an invasion of privacy. In the case of observation, however, if you inform all the users and seek their permission in advance, experiments show that their behavior is affected and the results skewed. The evaluation team will have to address this issue and decide on the most appropriate way of notifying users or addressing the problem in general. Having policies in place to ensure that data collected are used sensitively and anonymously can also help to address some of the ethical concerns. In addition, evaluators need to be aware of the privacy protections afforded by federal and state laws.

Evaluation can help us realize the full potential of the technology to create powerful and attractive applications that assist users to understand digital resources in meaningful and relevant ways.

Who should conduct evaluations?

A final question is who should carry out the evaluation work itself. Evaluation work requires time, resources, and skilled staff. It can either be conducted in-house or contracted to external professional survey or marketing companies, as is often the case with quantitative work, or to evaluation experts. Each option has different implications, advantages and disadvantages.

Carrying out the evaluation in-house can help to reduce costs, ensure that the work is carried out by staff members who are familiar with the mission of the organization, control the design of the evaluation and the way it is conducted, and provide for continuity between the evaluation and the way the organization takes advantage of the results. Before beginning, it is essential to ensure that staff have adequate training and expertise in designing and carrying out evaluation work and analyzing the data that are collected. For example, quantitative evaluation might require training in sample design and statistical analysis, while training in moderating focus group discussions might be useful for qualitative work, and system design skills may be essential for developing computer-based evaluation systems. Libraries, archives and museums may collaborate with specialists in information studies departments or with assessment professionals at area universities or colleges who can provide them with access to the necessary skills. One difficulty is that evaluation work is very time-consuming, and using in-house staff to carry it out will inevitably divert them from other activities. If you lack the skills in-house then it will prove more cost-effective to outsource the activity. For large-scale studies, that may entail drafting an RFP (Request for Proposals). The RFP should scope the problem that you hope to address through evaluation and ask the respondents to outline the methods they would use to conduct the evaluation work and how they would propose presenting the results.

Conclusion

Evaluation, like project documentation, is often viewed as ancillary to the main project goals, a separate task which takes a back seat to the work of creating the digital resource itself and disseminating it to users. However, as this section has suggested, evaluation not only offers tangible benefits in improved design and better responsiveness to user needs, but also may help avoid disastrous missteps in the planning stages, while change is still possible. Evaluation also enables you to document your project's effectiveness and make a persuasive case for new or continued funding. Careful planning will ensure that evaluation does not get in the way of efficient project development, or consume a disproportionate share of project resources. Conducted in this way, evaluation is an extremely valuable component of digitization work.

---

[1] For example, see http://www.ics.uci.edu/pub/websoft/wwwstat/, http://www.extreme-dm.com/tracking/, or http://www.analog.cx/

XIII. Digital Asset Management

Introduction

The *Guide* has already detailed the process of creating and distributing digital collections. But the process of digitizing material does not merely create an intellectual object: it creates a valuable asset, and consumes considerable resources in doing so. This section will look at mechanisms by which the institution that created or holds these digital assets can manage them to maximum advantage. It explores the value of the digital products to the organization, how they should be treated and managed in order to maximize their use, and how the institution can ensure that it receives full benefits from the sizeable effort and costs consumed in their creation. It describes the digital management options and assesses the need for digital asset management structure.

The digitization process produces digital assets whose significance is equal to any other asset that an organization may hold, such as computers, manpower or intellectual content. Indeed, they have been described as 'working' or intellectual capital. Just as an organization seeks to make efficient and effective use of its financial, manpower, and natural resources, it will now wish to use its digital assets to their full potential without reducing their value. In the digitization initiatives that were surveyed as part of our research, the digital creations will be presented and offered to the end-users in the same way as any other commodity is made available to customers. However, when the digital product is taken or 'bought', the organization still retains the product itself and does not need to engage in further manufacturing to reproduce the product. If the organization manages the product, for example through licensing (see Section IV on Rights), once the customer has purchased the digital material, it effectively remains an asset that the holding organization can exploit in perpetuity, provided that the value of the digital asset is not compromised.

The need for digital asset management

As the size of digital collections grows, so do the financial resources that are consumed in their creation. The costs of managing the resources also increase, although not necessarily in direct relation to the growth of the collection. It therefore becomes imperative to manage the resources effectively, a process that is commonly referred to as Digital Asset Management (DAM). There are many reasons for initiating the use of DAM.

Digital resources are often viewed as ephemeral and fragile, while at the same time they are seen as objects that can be easily recreated — one only need scan the picture again. Fortunately this lackadaisical attitude appears to be diminishing in most organizations and in its place is the growing belief that digital resources are, at the very least, as valuable as the time, effort, and finance that has gone into their creation. At the Berkeley

Art Museum, Günther Waibel considers the move from taking transparent film surrogates of the Museum's collections to taking surrogates with a digital camera to be a critical one. The digital objects become 'mission critical institutional assets' that require management strategies similar to those already in place for the Museum's existing collections (Waibel 2000). Digital Asset Management allows organizations to maximize the use of these resources, ensuring that their value is maintained, while generating institutional savings.

Digital Asset Management (DAM) involves:

- Creating an efficient archive that can hold digital resources (such as images, audio and text) and the metadata that describe them;
- Implementing an infrastructure to ensure that these electronic data are managed and preserved in such a fashion that they will not become obsolete;
- Implementing search facilities that enable users to identify, locate and retrieve a digital object.

The benefits of implementing DAM include:

- Centralizing discovery and access;
- Coordinating disparate projects as part of a coherent whole;
- Centralizing authorization, security, and tracking systems;
- Unifying organizational solutions to managing copyright and IPR;
- Reducing duplication of effort and resources;
- Saving time for the creators and users through organizational structure and centralization of data.

Digital Asset Management Systems

DAM systems provide the means to manage digital assets from creation to publication and archiving. In some cases, systems can automatically take the data from the scanning, assign the image a network space, depending on the metadata that the creator assigns to it, and then store the digital object and metadata in a database. A DAM system may be as simple as a directory of files on a hard disk, each file containing a digital asset, with an accompanying database that stores descriptive and administrative metadata for each of the files. Each database record contains metadata that can be used to find and understand the asset, including the name of the file and probably information about its content, format, history and usage. A simple asset management system can be purpose-built from

an off-the-shelf database management system, such as Filemaker Pro, Microsoft Access or one of the larger SQL database systems, like MySQL or Oracle.

Many cultural heritage institutions have computer systems that automate traditional functions. For instance, museums often have collection management systems for managing records of their artifact collections internally. Libraries usually have an online public access catalog for patrons searching the collection. DAM systems are a new addition to this family of computer systems. Once an institution has created a repository of reusable digital content (images of artifacts, video of public talks, etc), it will need to manage this new resource as a type of documentation collection, but one with special management needs. DAMs help the institution to manage these digital assets, and include information about the original artifact or event that the digital content relates to, as well as technical information about the digital resource, intellectual property rights to the digital resource (not the original artifact), and other types of relevant metatada that enable the institution to preserve and re-use these digital assets. Such a system can be used to administer a collection of assets, ensuring that each can be found and used by searching the data and locating the information about how to find the file. Usually, the system is intended to provide access for more than just simple management of the assets themselves, such as providing public access to the collection through a web site or using the assets to support the day-to-day activities of a museum or archive. However, delivery and asset management systems can be separate. It is also important to remember that a complete digital asset management strategy must start with the creation or acquisition of the assets. Tools that support the preparation of the data, particularly for processing batches of assets, are a very important part of the package.

In practice, such systems are almost always more complex than the simple one described above. There is usually more than one version of an asset to be tracked. For example, a primary image that is either digitized from a photograph or created directly with a digital camera may be processed manually to create a collection master; that master can then be mechanically manipulated to create derivative assets for specific purposes, such as a thumbnail and screen-sized version of the image for delivery on a web site. The result in this scenario is that there are four assets that all share an image content description, but each has its own administrative and technical description, and each may have different restrictions on its use. This example can be further complicated by considering a collection of digital images of objects or sites. There may be multiple photographs of the same object or site, each of which has a family of digital assets derived from it; the photographs share some description of the object or place and each photograph has some specific content description, all of which is inherited by each of the appropriate derivatives. While it is possible to exploit the relational database systems mentioned above to build a DAM system to handle more complicated situations, the technical expertise required is significant.

Unless the uniqueness of the project at hand requires a custom solution, it is probably better to use software already developed for digital asset management where possible, whether by purchasing a proprietary system or reusing a system developed at another institution. There are a number of proprietary products on the market that can be some

part of a DAM strategy. Unfortunately, at this point there are no obvious products that stand out as a unified, off-the-shelf solution to the general problem, though there are products that are appropriate for specific kinds of collections. The best advice that can be given at this time is for an institution to analyze its priorities, collections and intended audiences in order to develop a checklist that can be used to evaluate the products available and choose among the ones that seem appropriate. Some DAM tools have also developed by the cultural heritage community itself; for instance, a tool has been developed at Berkeley which has been used by several museums in the MOAC project (http://www.bampfa.berkeley.edu/moac/imaging/index.html).

Library software companies and vendors that specialize in museum collections management systems are beginning to extend existing products and develop new ones that can be a part of a DAM strategy. These products tend to be large-scale, expensive systems, with vendors usually working closely with clients to develop complete "solutions." In general, this path is probably most useful when considering DAM as a part of a complete collections management strategy. Terms to look for when looking at the systems include collections management, content management, and digital libraries.

Many more software products that are appropriate for DAM have been developed with large-scale web site management in mind. These systems are available for a wide range of prices, with widely varying functionality. Some systems are more specifically oriented towards text or image-intensive collections, while others can handle a wider variety of media. Terms to look for when searching for these systems include digital asset management, asset repositories, media catalogs, document management and content management systems. Note that the term "content management" is used very loosely, often to refer to software that is intended for administrative uses which would not be appropriate for DAM. However, note that some content management systems do include both DAM and delivery functionality (Emery 2002).

Developing a DAM strategy

Rarely is developing a digital asset strategy as simple as picking out which software package to buy and then implementing it with local constraints. It is unlikely that the needs of a project will be a perfect match for any software package; there will almost always need to be some customization to match the system with local requirements. It is either a matter of paying the vendor more for the customization of a system that does most of the job, or of buying or developing the additional tools needed. Unless the institutional context for the project includes strong technical support, a custom designed system will probably cost more than buying existing tools for the job. Even with good technical support, it pays to look closely before setting out to develop a system locally.

The key is to plan carefully, starting with the acquisition and creation of assets, looking at every process that leads to their ultimate uses and safe storage. Develop a checklist that addresses the five areas discussed below. What are the staffing implications that arise

from each of these areas? What are the data inputs and outputs at each step? What are the access points to the data, both within the institution and to the public?

File management

The most basic DAM issue is the management of the files that contain or constitute the assets. These files must have unique names so that they can be unambiguously addressed. Some proprietary systems handle all of the file storage for the user. File names are assigned by the system and the whole file storage array is essentially opaque to the user. This certainly can be useful in that it frees the user from the worry of tracking each individual file and makes it easier to control access to the digital assets appropriately. The disadvantage can come when trying to recover from a disaster, when upgrading a system or moving from one system to another; the user is completely dependent upon the software for access to the files. Operating system utilities necessary to these processes may be unable to address the files. It is best to establish what restrictions a software package imposes on file management with the vendor before purchase.

In more open proprietary systems, and in custom systems constructed around databases, the issues of file management are a very important part of the DAM strategy that should be addressed thoroughly in the early stages of planning. Strategies for naming files that seem intuitively obvious early in the process may become a problem as the project scales up. Naming schemes should be systematic, exploiting directory names, file prefixes and suffixes. It may be helpful to exploit existing identifiers that are known to be unique, like an accession number for museum objects. For a collection of digital images, the accession number could be used as the file prefix (appending a “v1”,“v2”, etc., for multiple views of the same object), the file suffix could reflect the encoding type of the each image, and each type of derivative could be stored in a separate subdirectory.

Another major issue for file management is creating backup copies of the assets that can be reloaded in the event of mechanical failure or operator error. This is most often done using magnetic tape. Though a detailed treatment of this subject is beyond the scope of this *Guide*, some basic principles can be pointed out. Backup strategies are usually a combination of making copies of new or changed files frequently and making full copies of the file system periodically. Tapes should be reused in rotation, retaining certain copies for longer periods, in order to be able to ensure an acceptable risk of loss. For example a busy project may make daily backups of files that change, with a weekly full dump of the file system. The daily tapes are then reused, with the weekly tapes being retained until a defined period has passed, at which point one of them can be retained for a longer period while the rest go back into the rotation. Many such schemes are possible and need to be investigated with respect to the specific needs of the project. Note that backup tapes also become assets that need to be managed. See Section XIV on Preservation for issues related to long-term retention of copies of digital assets.

Metadata definition and management

Metadata about digital assets are generally classified into three categories, all of which are useful in the management process: ***descriptive metadata***, which is about the content and form of the digital asset to enable search and retrieval; ***administrative metadata,*** which is about the history of the asset and policies associated with it, typically information on creation, quality control, rights and preservation; and ***structural metadata,*** which records information about the internal structure and relationship of resources to facilitate their navigation and presentation.

Any DAM effort must consider what metadata needs to be captured for digital assets early in the planning process. At this stage of technology (and for the foreseeable future) digital assets are pretty much unusable without metadata. The specifics of metadata for the different media are covered in other sections of the *Guide* and in the appendix on metadata, but there are some general principles that should be followed when planning a project.

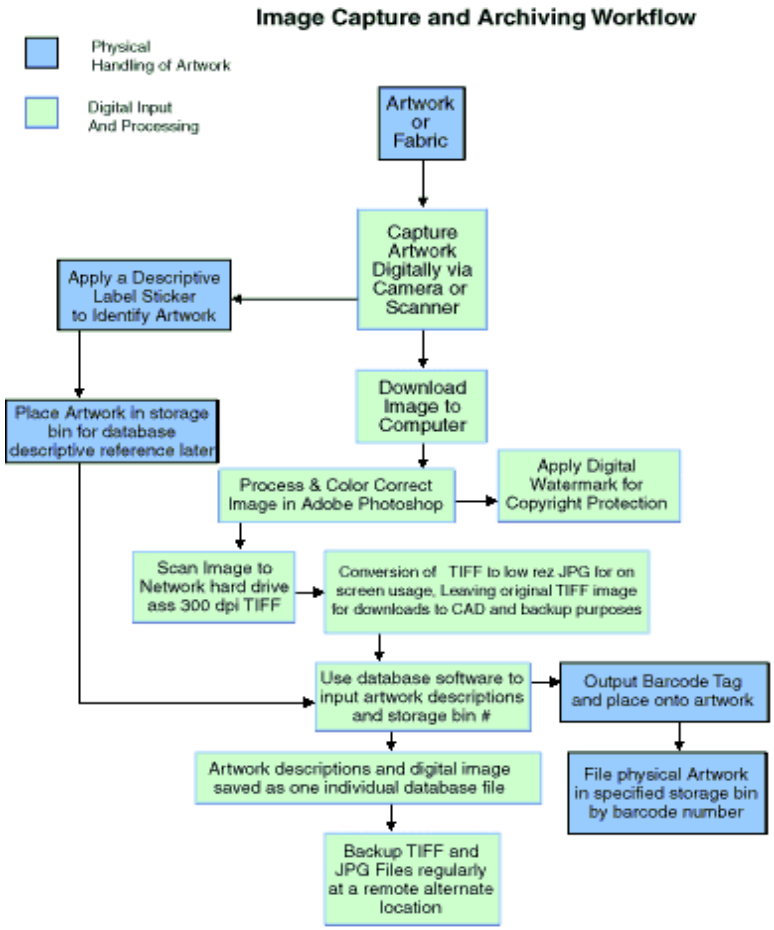
The effort associated with the creation of metadata can be equal to or greater than the effort to create the assets themselves, inasmuch as it represents an explicit statement of all the most important information about the asset. It is necessary to carefully consider the ways that the assets are intended to be used, who would be using them, and any legal or administrative requirements that are involved. Then the appropriate metadata must be captured or created to support all of the activities that require the assets, in addition to the file management activities described above and the workflow and access issues described below.

Because the DAM process is largely a function of managing metadata, careful attention should be paid to developing support for the creation process and for searching and reporting functions. If possible the system should be able to incorporate controlled lists of values (where appropriate) for metadata fields, either through user-defined lists or by incorporating standard lists, like thesauri and community-defined authority lists. The ability to search and to prepare reports from the metadata are both key functions in the DAM workflow.

Workflow

Any strategy for managing digital assets should address every step of the process, starting with the creation or acquisition of assets and metadata, through the processes of their day-to-day management, to delivery. Figure 1 shows the workflow for a process of capturing and archiving images, developed by Teri Ross (used by permission). This diagram illustrates that the DAM process can consist of many steps, requiring a variety of processes and possibly a number of different tools and skills. The development of a good digital asset management strategy should include an analysis of the necessary workflow in the planning stages of the project.

In any project the work of creating and acquiring both the assets and the metadata, and of migrating assets to new formats as necessary, can be substantial. Tools and utilities that make all of these processes more efficient and more reliable are very desirable. In particular, utilities that support batch processing of assets and the capture of metadata from the creation process are both very important parts of a DAM strategy. Note that many proprietary systems may not include all of the tools and utilities necessary to a good workflow so it is important to make sure that other software, purchased or created, can easily interface with the DAM system.



From Teri Ross, 'Digital Asset Management: The Art of Archiving', http://techexchange.com/thelibrary/DAM.html

Policy tracking and enforcement

A digital asset is only valuable to an institution if its use can be effectively managed. As is evident in other sections of this *Guide*, managing the rights associated with digital objects is complex but DAM systems can help resolve some of those complexities. Three

areas where DAM systems can help manage rights include assertion, protection and management. Protection comes in various forms, from managing access to the digital repository, to tracking users, controlling what versions of material users can access, and ensuring that rights metadata is linked to the object when it is delivered to the user.

Data relevant to other types of policies may be a part of a DAM strategy. If the assets produce revenue for the asset holder, the DAM system must at least be able to interact with the fulfillment process. Versioning of assets is another area that can complicate a system. If new versions of the asset are created over time and all versions remain available to the user, it may be important to track the different versions and deliver a specific version on demand.

Policy tracking consists mostly of keeping the appropriate metadata about assets, and, in general, it is a fairly easy process to manage. The basic metadata management of a DAM system described above should be adequate. The systematic enforcement of policies through software is very much a research topic rather than a feature of systems that are available. It is reasonable to expect that systems should be integrated with the relatively simple access control that is available through the current generation of web servers. More sophisticated needs for policy enforcement should be discussed with vendors and/or developers with an eye to the long-term feasibility of the DAM strategy rather than as a necessary feature.

Access

In addition to their availability for public access, which is covered in the Distribution section of the *Guide*, digital assets are often needed in the day-to-day processes of the institution that owns them, possibly requiring the DAM system to interface with other systems. For example, a museum's exhibition design group may use images of artworks from the general collection in planning exhibitions. Their process would need to be able to access those images and their metadata.

In elaborate DAM systems, this kind of access is often part of the package. But even there, careful consideration should be given to the details of how the data is made available. Complete systems are built upon assumptions that may be overly restrictive, making it difficult to adapt to new demands in the future. In general, a DAM system should be able to accept input and deliver output in standard streams. The availability of good, open application program interfaces (APIs) with the system can ensure that interoperability can at least be developed.

Conclusion

Digital asset management at its most basic level uses technology, such as commercial off-the-shelf applications, to implement common sense goals of organizing resources, enabling users to discover them and owners to track them. At its most advanced level, a DAM system can radically transform the way an institution manages digitization and

handles access to the digital assets. It is important to develop a complete strategy that covers the complete continuum from creation to delivery: storing the digital objects and their related data; supporting on-line searches; enabling users to retrieve the appropriate version of a particular digital resource depending upon their interest and/or access level; keeping track of rights restrictions; and limiting access to verified users. DAM makes workflow more efficient and can assist organizations save money and time. Digital objects are valuable assets for projects and institutions, and DAM systems offer a rich array of ways to ensure that these assets retain their value and are properly managed.

XIV. Preservation

Introduction

The projects and programs interviewed for this *Guide* all acknowledge that digital materials are at risk of loss, both through physical damage to the storage medium and through technological changes that render the data unreadable. Although none of the projects we interviewed identified digitization as a chosen method for preserving analog material, digitization is starting to supplant microfilm as a preservation medium and hence the urgency of providing long-term preservation of digital materials becomes of paramount importance. For museums and institutions with two- and three-dimensional objects, moving images, and audio, the need is even greater, since microfilm has never been even a minimally adequate preservation medium for these. Overall, there is an urgent need for significant research and the development of affordable strategies for trustworthy digital preservation services. This *Guide* does not aim to cover the full range of technical and scientific developments in storage media and methodology, but explores some of the issues involved in preservation in order to explain the practical steps you can take to store and preserve your digital materials.

In this *Guide* we define 'preservation' as the actions taken to preserve digital objects, whether they are created from analog collections or are 'born digital' objects such as original digital art, and digital audio or video that are not re-recorded from analog originals. This domain includes two important and overlapping areas: first, the goal of presenting the content of analog items via digital reformatting; and second, preserving digital content regardless of its original form. 'Conservation' of the original analog objects—generally the domain of conservators, museum curators, librarians, and collection managers—is somewhat different, as it involves the care and repair required to keep the physical object in existence. However, there is some overlap between the two: objects are often digitized to help conserve the original object by reducing wear and tear on fragile and vulnerable materials, while still allowing access to them.

Preservation of digital objects can be thought of as long term responsibility to the digital file — responsibility to ensure that the file retains the information it had when it was created. Keeping the content 'alive' over time entails much more than simply keeping the files intact. Issues such as metadata must be understood as well. These issues are explored more fully in the rest of the section. There are two distinct areas of intention here, although in practice they often merge: preserving the digital surrogate, so that its content remains accessible and usable over the long term; and using digitization as a method of preserving objects which are valuable, fragile or transient. For example:

- Access to digital surrogates replaces direct use of rare and fragile originals, aiding their conservation.

- Some projects use digitization to 'capture' objects that are discarded after the capture process, such as newspapers.

Therefore the digital surrogate is a form of preservation (although not a substitute for any other form of preservation) and must itself be preserved to ensure future access and use (see Section XI on sustainability). The project should consider the two areas: digital reformatting and preserving content in digital form.

Why preserve?

Digital preservation is an essential aspect of all digital projects: it is imperative that the digital objects created remain accessible for as long as possible both to intended users and the wider community. Decisions made throughout the digitization process will affect the long-term preservation of the digital content. While the processes involved in digital preservation are the same whether for digital surrogates or for born-digital objects, remember that born-digital objects are the only version of an artifact or object so are, in effect, the preservation copy. All these factors must be incorporated in your digital preservation strategies.

CEDARS (http://www.leeds.ac.uk/cedars/) has described the digital preservation process as follows:

> *Digital preservation is a process by which digital data is preserved in digital form in order to ensure the usability, durability and intellectual integrity of the information contained therein. A more precise definition is: the storage, maintenance, and accessibility of a digital object over the long term, usually as a consequence of applying one or more digital preservation strategies. These strategies may include technology preservation, technology emulation or data migration. There is a growing wealth of information on digital preservation and related issues available on the Web.*

The term 'accessibility' as used above deserves further glossing. It is important to distinguish between what we might term 'machine accessibility', involving the data integrity of the original files and digital objects, and human accessibility, or the ability to render the object intelligible and usable by a human reader. The latter is the more complex challenge, since it requires not only that all the original bits be present in the right order, but also that there be software available that can interpret them and display them appropriately.

Digitization projects are costly and time consuming, so it is essential that the digitization process should not need to be repeated, incurring more expense and subjecting the originals to further stress. The process may not even be repeatable: some projects we interviewed do not keep the original object. The Library of Congress, for example, converts newspapers to microfilm in certain cases, discarding them when the process is complete. The University of Michigan and Cornell have both discarded original objects for a variety of reasons, such as maximizing shelf space. Now the digital files are the *de*

facto preservation copy, so it is even more important that they be suitable for long-term use and remain accessible for future generations.

Preservation should not be considered as an 'add-on' to the whole digitization process; the decisions taken throughout — choosing file formats, metadata, and storage — should all consider the long-term sustainability of the digital objects.

Preservation and storage of the digital surrogates

The projects surveyed saw digitization not as a way of creating preservation surrogates from original analog materials, but rather as a reformatting exercise designed to enhance access to and usability of materials. However, as noted above, projects like those at the University of Michigan, Cornell, or the Library of Congress, which discard the analog originals following digitization, must treat the digital version as a preservation surrogate. Furthermore, any project dealing with unstable materials must have similar concerns, since their digital version may in time become the only record of the original. Nonetheless, all of the projects interviewed recognized that they were creating assets that need to be preserved in their own right. Ensuring their long-term availability was particularly important where scholarly work depended upon or referenced them. As the limitations of microfilm become more apparent, many in the cultural heritage community are recognizing digitization as a means of preservation of digital content as well as improving access. Remember that here we are not so much concerned with conservation of the original but with preservation of the digital surrogate itself.

The starting point in planning digital preservation is to identify what can actually be preserved. Various levels of digital preservation are possible, from the 'bucket of bits' or bitstore (the storage of files in their original form with no plan to further reformat them for different forms of accessibility or searchability) through to preserving the full searchable functionality of an image or audio system. Each project must decide what is crucial, manageable and affordable. Consider a range of options, from the absolute minimum up to the ideal. If you decide to preserve the minimum, or the bucket of bits, then migration of software and operational systems will be less of a concern. This option might be attractive if your goal is simply to preserve archived data created by others, where your role is merely to keep the data, not to provide an ongoing framework for active use. However, if you decide to preserve the whole system you will need further migration of data to ensure usability in the future. The basic minimum will depend on each project, its rationale and its long-term user requirements. There is no limit to what can be preserved in digital format, apart from the limitations imposed by migration, storage space and other resources, such as staff time.

There are four main issues to consider in formulating your preservation strategy.

The first is software/hardware migration. All products of digital preservation must be migrated at some point, at the very least to a file format that the latest technology can recognize. If you have chosen to preserve the whole system, then operating systems and functional software must be migrated as well. This can cause problems as upward

compatibility is notoriously uncertain, even from one version of software to the next, and there is no guarantee that the version you are using will be compatible with releases in many years to come. Full system migration must be carried out frequently to ensure access and usability. You will need to formulate a migration policy that is implemented on a regular basis rather than as a reaction to new software or hardware. Regular testing after migration is also crucial to ensure that functionality has been preserved. This is conventionally called system emulation.

The second issue concerns the physical deterioration of digital media. All digital media deteriorate over time, but this process will be more rapid if they are stored in an inappropriate way, such as in a damp basement, or as a pile of CDs stacked one on top of another. Correct storage (e.g. in racks that enable the disks to be stored separately) and an environmentally controlled location will help to optimize their stability and protect them from loss. The Genealogical Society of Utah, for example, stores the archival digital master files in an underground vault in Utah as well as in another state in the USA. Tape storage is useful, as there is strong evidence for the extended lifetime of this medium when stored correctly; remember that all tapes must be spooled regularly to avoid sticking. As with software and hardware migration, digital media should be checked and refreshed regularly to ensure that the data are still readable, and this process should be part of your preservation policy. All institutions should draw up a policy that reflects the value of the digital material and therefore sets out how regularly the media should be checked and replaced. This aspect of the preservation policy is linked to hardware migration, as some media formats may no longer be readable by the latest computer hardware. Preserve your data on a medium where the hardware exists to transfer to later media if the original becomes obsolete. Remember that it is costly to use a data recovery agent to move files from an obsolete medium, so make sure your preservation policy will prevent this happening, and migrate while the process is still straightforward.

The third issue concerns metadata, which is crucial to preservation of digital resources. The level of metadata recorded and stored will, once again, depend upon what you choose to preserve, the minimum content or the fully functional system. The format of the metadata is also important, as the metadata should be as accessible as the data. ASCII will not need to be migrated, so it is the most durable format for metadata, although it lacks the functionality that a database might give. SGML/XML files (including HTML) are stored in ASCII format and provide a high level of structure and functionality without requiring proprietary software. The site structure and the relationships between the files should be recorded with a METS format XML file, and the other kinds of files (e.g. images, databases) that make up the site should, if possible, have preservation copies made. Word-processing files should be copied as ASCII text files, and if possible converted to XML to preserve their structure. All databases and Excel spreadsheets should be exported as comma-delimited or tab-delimited ASCII text files. PowerPoint files should be exported to HTML with gifs, with those gifs potentially re-saved as separate TIFFs. GIS datasets should be exported as comma-delimited ASCII.

A fourth and more complex issue is the question of user needs and preferences, which may cause certain formats to become effectively obsolete even while they remain

technically functional. For instance, user discontent with microform readers threatens to make microfilm obsolete even though it still fulfills the original goals of its creation. User acceptance—and its decline—will be one of the key "trigger events" that will compel migrations to new delivery versions of digital collections.

Technical, media and file formats

Careful consideration of the technical issues will help you to ensure the long-term accessibility of your digital material.

As outlined above, the physical media on which the digital material is stored need to be managed in monitored, controlled systems and environments to avoid physical deterioration through human mishandling and inappropriate storage. Magnetic media such as digital audio and video tape are highly vulnerable to fluctuations in temperature and humidity, and of course to magnetic fields. Optical media such as CD-ROMs and DVD-ROMs are more durable, but should still be stored with care. For all media, it is advisable where possible to store a second full set in an off-site location, to guard against the risk of theft, fire, and similar disasters.

Choice of medium is equally important. Formats change rapidly and obsolescence is a perennial problem; it may be very difficult in the future to find a drive that can read ZIP disks, just as now it may be difficult to find readers for certain kinds of magnetic tapes. No one format can be guaranteed to persist and remain accessible in years to come, so you need to anticipate continuing developments and new products. Among currently available media, DVD, DLT (Digital Linear Tape) tape and CD-ROM are the most popular. Tape has a tried and tested longevity, while DVD and CD-ROM are newer media and current research has not yet agreed on their likely lifespan. Best practice in preservation is to migrate data from one medium to another, for example from optical disc to tape, while the hardware and software are still available. Migration is an integral part of any digital project and should be set out in your preservation policy.

If projects choose to make back-up or archive copies on fixed media, then it is crucial to ensure that the fixed media is stored correctly and that migration to the latest type (e.g. DVD) is carried out regularly.

A similar obsolescence problem will have to be addressed with the file formats and compression techniques you choose. Do not rely on proprietary file formats and compression techniques, which may not be supported in the future as the companies which produce them merge, go out of business or move on to new products. In the cultural heritage community, the *de facto* standard formats are uncompressed TIFF for images and PDF, ACSII (SGML/XML markup) and RTF for text. Migration to future versions of these formats is likely to be well-supported, since they are so widely used. Digital objects for preservation should not be stored in compressed or encrypted formats.

While TIFF and PDF are *de facto* standards, it must be remembered that they are proprietary formats with multiple 'flavors.' Not all image software can represent TIFF

files accurately and Acrobat is required to view PDF files. Projects should monitor the changes in these formats and ensure that users will be able to access the files over time, either by migration or emulation.

Storage and retrieval

The readability of your files is dependent on application software and operating systems. Consideration of file format selection must take this into account. Access software, operating systems and user interfaces are subject to the same process of continuing development and change as storage media and file formats. New versions of operating systems will affect the usability of your files, even though the file format itself may still be accessible. Remember that your operating system may well be proprietary software; unless you are using a non-proprietary system such as Unix, you will have to purchase the more recent version and migrate the files.

It is advisable to separate the underlying content from the discovery and display systems.

Migration will be more complex for those projects that have chosen to preserve the complete integrated system, with full searchable functionality, as there are more features to capture and preserve. Backwards compatibility is often limited and so, as with media migration, these programs should be monitored closely to enable migration before obsolescence makes this impossible. For example, you may be using proprietary software which has regular (often yearly) upgrades. The versions may eventually change so much that backwards compatibility is no longer viable, resulting in, at best, a long and expensive process of data extraction and migration, and at worst, the loss of the data altogether. To ensure accessibility, migrate every file to every new version. The use of standard data formats such as SGML and XML can be an advantage here, since it keeps the data independent from the delivery software. Even if the latter becomes unsupported or if cost forces you to switch to a different delivery system, your data will remain intact and portable.

Link Box:

Preservation of Electronic Scholarly Journals: http://www.diglib.org/preserve/presjour.html

This project, a collaboration between CLIR, CNI and DLF working with the Andrew Mellon Foundation, aims to develop plans for affordable preservation strategies that will implement the OAIS reference model.

Monitor the user interfaces, too. Scripting languages, plug-ins, and applets may all become obsolete over time and users will not be able to access your materials.

In general, remember that standards are the key to creating a resource that is interoperable (the ability of the resource to be used by a different system) as well as sustainable over time. Relying on proprietary software or formats that are not accepted

community standards may mean that you will have to migrate more frequently and in isolation. All projects will have to migrate their digital files at times, but by using standards this process will be less frequent and may take place when all projects are migrating data in response to a general change, such as that from SGML to XML, making it easier for everyone.

Organizational questions

The long-term view taken by institutions will also influence the sustainability of digital objects. Projects that have established preservation policies and procedures, such as Cornell and the Online Archive of California, are identifying preservation issues and placing them at the heart of their institution's ongoing procedures. Insufficient institutional commitment to long-term preservation can create digital resources with limited sustainability.

Similarly, human and financial resources play a huge role in implementing such policies and procedures. Staff turnover and variations in usage will influence the level of commitment to individual projects. Unsatisfactory record keeping and metadata, administrative as well as preservation and access, can also contribute to institutional neglect of digital materials. Moving from individual projects to organizational digital programs will facilitate the long-term preservation of digital materials, by providing greater organizational stability and the overhead necessary to maintain good records and documentation.

Addressing the issues

Whether you have opted to preserve the minimum content or the whole discovery and display system, policies should be in place to ensure the long-term sustainability and accessibility of the digital files you have chosen to be preserved.

We found that attitudes to digital preservation vary enormously. Many of the major digitization projects are using research sources, such as D-Lib Magazine (http://www.dlib.org/), to help them understand the trends and strategies in digital preservation. At the same time, we found that others did not know what digital preservation was and were uncertain of its importance for their projects. There are many questions and no single answer to the preservation issue. No one project is satisfied that it has found the best storage and migration policy to ensure longevity. The Genealogical Society of Utah project is taking a larger view of the issues through its research with Brigham Young University to identify storage media and standards that will help ensure longevity of digital surrogates. All the projects agreed that adopting standards in creating digital objects would enable their sustainability. Each project will choose what it will preserve based on the level of resources available for storage and migration as well as user needs.

Research is being carried out that will help you make the right decisions for your digital resources. Projects, such as CEDARS, have attempted to establish strategies for ensuring digital preservation through technology preservation, technology emulation and data migration. VidiPax and other companies are addressing the technical media side through research into restoring and archiving digital media, while organizations such as NARA are addressing both traditional concerns (records management) and new technical standards. Internationally, The Digital Preservation Coalition in the UK is working to foster joint action to address the urgent challenges of securing the preservation of digital resources in the UK and to work with others internationally (http://www.jisc.ac.uk/dner/preservation/prescoalition.html), and Preserving Access to Digital Information (PADI), a project of the National Library of Australia, maintains a useful web site with international research results and recommendations (see Link Box below).

The key to understanding and implementing digital preservation is to view it as a series of tasks rather than as one process. These tasks, if properly carried out, will help ensure not only that digital objects are correctly stored but also adequately maintained and useable over time.

Although there are others, six methods of digital preservation are widely discussed among heritage professionals. They are not mutually exclusive, and you may decide on one or a mixture of methods.

Technology preservation relies on preserving the technical environment in which the digital objects were created, both hardware and software. This approach is more relevant to projects that have chosen to preserve the full system with functionality and interface as well as the actual content. The digital files must be preserved as well as the computer and software. Maintaining hardware for long-term use is problematic, and it is also costly to store the physical machinery as well as the digital files. This method would apply to projects that wish to maintain the environment for historical use (a computer museum, for example) and may not be appropriate for many cultural heritage projects.

Technology emulation also seeks to preserve the technical environment, but instead emulates the environment on current technology, mimicking the software and even the hardware. Everything has to be emulated, including operating systems, scripts, and applications. This is unlikely to be a viable approach for most digital imaging projects. HATII's Digital Archaeology Report for the British Library (http://www.hatii.arts.gla.ac.uk/Projects/BrLibrary/index.html) has more detailed explanations of emulation.

Data migration focuses on maintaining the objects in current format and may involve migrating all files to a newer format (for example in 2001, from JPEG to JPEG 2000), when current software no longer supports the earlier file format. Data migration is a viable option for cultural heritage programs and will involve making certain decisions, such as how often to migrate and what format to migrate to. It can be a time-consuming process, involving inevitable loss of data as well as sacrificing an element of the 'look

and feel' of the original material. Using community standards at least gives the reassurance that every project will be migrating data of similar types at roughly the same point in time, making it more likely that tools and support will be available. The Archivo de Indias project has demonstrated that it is possible to migrate millions of documents form one file format to another, provided you design the work program effectively.

Enduring care encompasses all aspects of caring for the digital objects and the media they are stored on, including housing in secure locations and careful handling. This is short-term care and projects must consider migration and refreshing for longer-term preservation.

Refreshing is the process of moving data from one medium to another, e.g. from CD-R to DVD. It is not a total guard against obsolescence and should be seen as part of the whole digital preservation strategy.

'Digital Archaeology' describes what to do in a 'worst case' scenario, where a digital preservation policy has not been followed, or an unforeseen catastrophe has damaged the media. As described in the HATII report on Digital Archaeology, this is the process of rescuing content from damaged media or from obsolete or damaged hardware/software.

Storage Media

Data migration should not be confused with *'data refreshing'*. Data migration is the act of moving the actual content of the digital files to a newer file format (e.g. from Access 97 to Access 2000), while data refreshing is moving the files (no matter what format they are in) to a new physical medium, e.g. from CD to DLT tape, either because of obsolescence (e.g. laser discs) or because of physical deterioration.

Each project will have to identify the best storage strategy for its digital masters. This will include the choice of medium and the choice of institution/organization entrusted with managing that medium. The benefit of choosing a management service is that you can forego the responsibility and effort in choosing the medium. The benefit in investing the time to choose the medium is that you enjoy total control over your collection. Favored media vary from project to project and depend upon costs, access issues, physical storage space and technical equipment available. An in-depth examination of the technical issues of digital media is beyond the scope of this *Guide*. For a fuller explanation of these issues see the HATII report on Digital Archaeology cited above.

The formats used by the projects interviewed for this *Guide* include:

CD-ROM (including CD-R)

DAT Tape

DVD - both DVD - Video and DVD- ROM

DLT Tape

RAID Server

A mixture of the above media

Projects using audio-visual material will need to use a medium that can deal with the large file sizes created when digitizing such materials as well as enable access to the stream of data. DVD and DV Stream are the most popular media for these formats.

CD-ROM is a popular storage medium — it is cheap and simple to process. Projects such as the Online Archive of California, the Digital Imaging and Media Technology Initiative of the University of Illinois and the Colorado Digitization Project use CD to archive their digital materials. CD writers are easy either to install and run from the machine that is temporarily storing the data, or to back up from the server that holds the access data. CD-R has less storage capacity than DLT tapes or DVD — CD-R can hold approximately 640 MB of data while a DLT tape holds 35-70GB. Magnetic tape is known to be stable for approximately 30 years under good storage conditions. DLT tapes for instances should survive for a million head-passes over a 30 year period. Manufacturers claim that CD-Rs will last from 50 to 200 years, but experience indicates that even after as little as 10 years they are beginning to suffer from drop-outs. Media lifespans are an unknown quantity, so you may count on having to migrate and refresh the data regularly. These figures are based on climate- and temperature-controlled storage conditions and safe environments. Lifespan also assumes that the media are free from manufacturing errors and are not mishandled.

You can act to improve the storage life of your media. We recommend that you handle media with care and that you check it regularly. The following two tables give examples of how to improve the lifespan of various storage media.

Good Practice Box:

Improving the lifespan of CDs and DVDs:

Avoid	Never	Always
Damage to the upper and lower surfaces and edges of the disc	Attach or fix anything to the surface of the CDs	Store media in a jewel case or protective sleeve when not in use
Scratching and contact with surfaces that might result in grease deposits (e.g. human hands)	Write on any part of the disk other than the plastic area of the spindle	If using sleeves, use those that are of low-lint and acid-free archival quality
Exposing discs to direct sunlight		Wear gloves when handling the master CDs

Good Practice Box:

Improving the lifespan of DLTs:

Avoid	Never	Always
Placing the tapes near magnetic fields	Stack the tapes horizontally	Keep tape in its protective case when not in use
Moving the tapes about	Put adhesive labels on the top, side or bottom of cartridge	Move tapes in their cases
	Touch the surface of the tape	Store the tapes in appropriate
	Put a tape that has been dropped in a drive without first visually inspecting it to make certain that the tape has not been dislodged or moved	Store the tapes vertically

There are two types of DVD: DVD-Video and DVD-ROM, which can be compared respectively to the audio CD and the CD-ROM. DVD-Video can contain a whole film with digital audio. The DVD-ROM stores computer-readable data.

Many larger projects store two copies of the media, in different storage facilities. Both the GSU and SHOAH use underground vaults with temperature controls and high security in two geographically remote locations. SHOAH has the added risk factor of earthquakes and has chosen its second site in Washington DC. Not all projects can afford such facilities and the additional security that they provide, and will simply store one copy on site. We recommend, however, that all projects store archived master files off-site where possible to ensure against natural and human disasters.

We found that many projects did not check archived data for refreshing and did not have a preservation strategy at all, although many are developing them.

Stargazing Box:

There is a need for better long-term storage formats. The GSU project is conducting very interesting research into media for digital preservation. Through their work with BYU and Norsam (http://www.norsam.com/), they are examining media such as the HD-ROM, which has the following manufacturer's claims:

HD-ROM and the "ET" System are planned to have the following specification features: (ET is Electrical Transport Data Storage System)

- 200 Gigabytes per CD-sized disc
- 2 terabyte 10-disc Modular Reader
- 60 terabyte Modular Storage Unit which holds 30 10-disc Modular Readers
- Write rate of 30 megabytes per second
- Expandable to petabyte storage
- High-speed random access
- Each disc will have its own read head
- Read rate comparable to CDs

File formats

The file formats you choose are crucial to ensuring data migration. In 2001, uncompressed TIFF is the preferred raster image format for digital masters. Migration from this format may be inevitable to ensure long-term access to the materials. However, TIFF is widely supported and software manufacturers may continue to support it for the foreseeable future. It should be monitored carefully. TIFF is platform-independent. Archival copies should be stored uncompressed.

JPEG, JPEG 2000, MrSid, and GIF formats—which may be well suited for delivery—are not considered to be as well suited as TIFF for preserving digital master images, for two main reasons. All of these formats incorporate compression algorithms that offer potential benefits for storage and networked delivery, but increase the requirements for both monitoring and migration. Compression is one more thing to worry about when tracking the availability of software that can read and write formats without introducing information loss; thus, it is possible that intervals of migration will be more frequent for compressed formats rather than for uncompressed TIFF. In addition, because these formats employ lossy compression algorithms, there are opportunities for information loss at the time of capture and in subsequent migration activities. Other formats with lossless compression, such as PNG, are not widely used yet because of their relatively

limited support within software packages and image viewers. TIFF, while not an open standard in the strict sense, is the *de facto* standard format for digital masters within the community, but not necessarily for delivery, given the fact that standard web browsers will not display TIFF images.

Formats for audio-visual material include WAV, AIFF, MPEG (various versions) as preservation archival copies. RealMedia (both audio and visual) and QuickTime are used for streaming data to users but not as archival formats.

Encoding for longevity

The structure of digital objects is just as important as media and format in ensuring their longevity.

Digital objects stored in a proprietary format, such as a proprietary database, are difficult to migrate forward; proprietary software often has a short lifespan and is vulnerable to changes in the company that owns and supports it. A proprietary database can be unreadable within five years. You can minimize this risk by exporting your data in a standard format (such as tab-delimited ASCII). Similarly, use standard encoding schemes to structure your digital objects in order to minimize the risks of data obsolescence. Encoding standards such as SGML or XML are platform- and software-independent, and many projects, including some of the larger ones such as those at the Library of Congress, Cornell, Michigan, the Online Archive of California, and Virginia, have adopted these formats.

Metadata

One of the most important factors in ensuring digital preservation is the creation of robust metadata to describe the process of digitization as well as the actual digital objects themselves. Administrative metadata are particularly associated with digital preservation and the retrieval of the images from the archived media. This facilitates both short-term and long-term management and processing of digital collections and includes technical data on creation and quality control. Administrative metadata also handle issues such as rights management, access control and user requirements as well as preservation strategy information. The administrative metadata should include technical data that can be of potential use to systems (e.g., for machine processing tasks during migration), and to people, such as administrators monitoring the number and types of formats in the repository, or even to users inquiring about the provenance of a digital object. Having metadata in place will facilitate digital asset management, by helping to manage the digital files, their location and their content. Also see Section XIII on Digital Asset Management.

Link Box:

Some projects are adopting an institutional approach to metadata.

- Metadata Encoding and Transmission Standard (METS) http://www.loc.gov/standards/mets/
- CEDARS project: structure for preservation metadata: http://www.leeds.ac.uk/cedars/metadata.html
- Preserving Access to Digital Information (PADI): Research Overview and Updates on Preservation Metatada: http://www.nla.gov.au/padi/topics/32.html
- NISO: Technical Metadata for Digital Still Images: http://www.niso.org/standards/resources/Z39_87_trial_use.pdf
- OCLC/RLG Preservation Metadata Working Group: http://www.oclc.org/research/pmwg/
- Reference Model for an Open Archival Information System (OAIS): http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf
- Amy Friedlander, "The National Digital Information Infrastructure Preservation Program: Expectations, Realities, Choices and Progress to Date," D-Lib Magazine (April 2002) Volume 8 Number 4; http://www.dlib.org/dlib/april02/friedlander/04friedlander.html
- OCLC/RLG Preservation Metadata Working Group: http://www.oclc.org/research/pmwg/
- A Recommendation for Preservation Description Information: A Report by The OCLC/RLG Working Group on Preservation Metadata: http://www.oclc.org/research/pmwg/pres_desc_info.pdf
- Anne Kenney et al., "Preservation Risk Management for Web Resources: Virtual Remote Control in Cornell's Project Prism," D-Lib Magazine (January 2002), Volume 8 Number 1; http://www.dlib.org/dlib/january02/kenney/01kenney.html

While administrative metadata are crucial to the long-term usability of digital materials, structural and descriptive metadata are also critical to the digital preservation process. All projects recognize the importance of good, robust metadata and have implemented schemes that facilitate the maintenance and accessibility of the resources. The importance of high quality metadata cannot be underestimated, as this can ensure that while data migration occurs, the original information about the digital object is not lost and can be recreated if necessary.

Organizational strategies

Over the last few years, those involved in developing and delivering digital objects have realized that preservation issues go far beyond questions of preserving bits and dealing with obsolete media and file formats. Indeed, there is widespread recognition of the need to develop trusted digital *repositories* that can ensure not only migration of media and formats, but also the preservation of relevant information about digital objects so that we can preserve their behaviors as they are migrated forward in time and make meaningful use of them in the future. We need to preserve not just the byte stream, but also the structure of the digital object and its context. Discussions about the attributes of digital repositories have crystallized around the emerging ISO standard, "Reference Model for an Open Archival Information System," originally developed by NASA to deal with the huge volume of digital material developed by the space program. (In May of 2002, RLG-OCLC issued a report based upon this model entitled "Trusted Digital Repositories: Attributes and Responsibilities.") The Draft standard states:

> *An OAIS is an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community.... The model provides a framework for the understanding and increased awareness of archival concepts needed for long-term digital information preservation and access, and for describing and comparing architectures and operations of existing and future archives.*

OAIS is a reference model, not an implementation plan. Libraries and archives are just now beginning to design digital repositories. [1] As they look to implement OAIS systems, metadata concerns loom large. Long-term preservation of digital objects in a repository requires that we develop and maintain structural and administrative as well as descriptive metadata about those objects. The METS initiative has gained much attention this past year as a flexible way to address these needs. The FEDORA project, which is complementary to METS, is aiming to develop behavior metadata as well. At this time, creators of digital objects and projects do not yet have full guidance on what digital repositories will need to secure their objects for the long term. In particular, much of the progress to date has focused on ingesting, storing, and preserving individual digital objects and much more research needs to be done on the long-term needs of complex digital creations. Nonetheless, attention to the emerging structure of repositories and their metadata needs will definitely increase the likelihood and ease with which digital objects can be incorporated into repositories and preserved for long-term access.

One of the major issues involved in creating trusted repositories is that of responsibility. To what extent will creators and local providers need to cede control over their digital objects to centralized repositories? The advantages that large, financially secure institutions can provide in terms of long-term preservation are unquestionable. Nonetheless, research efforts especially amongst the digital library community are also focused on developing distributive systems. Indeed, there is a growing sense that "lots of copies" will be an important preservation and authentication strategy for the foreseeable

future. With the ever-declining cost of storage space, redundant distributive systems with shared responsibilities has many advantages that mitigate against centralized control.

Conclusion

Digital preservation is a discipline that is still in its infancy. Much of the preceding discussion provides a general description of various conceptual approaches to digital preservation, and also provides guidance for good custodianship (in the present) of existing digital files and digital media. However, it is expected that, in the near future, digital preservation will be executed within the context of enterprise-wide digital repository systems. This is an area of ongoing research and development. For example, in the United States, there is the new National Digital Information Infrastructure Preservation Program (Friedlander 2002), whereby the Library of Congress will lead a national planning effort for the long-term preservation of digital content and will work collaboratively with representatives of other federal, research, library, and business organizations.

Summary

Key factors in preservation are:

- Identifying the data to be preserved — basic content and metadata through to the fully functional program with interface look and feel
- Adopting standards for file formats
- Adopting standards for storage media
- Storing data on and off site in environmentally secure locations
- Migrating data
- Refreshing data
- Putting organizational policy and procedures in place

---

[1] Examples of projects using OAIS are CEDARS, NARA, NEDLIB and PANDORA. A chart for the deposit system for electronic publications can be found at: http://www.dlib.org/dlib/september99/vanderwerf/09vanderwerf.html. The CEDARS project aims to produce strategic frameworks for digital collection management policies, and to promote methods appropriate for long-term preservation of different classes of digital resources, including the creation of appropriate metadata.

Appendix A: Equipment

This appendix brings together material from various sections of the *Guide*, in expanded form, to provide a detailed account of the kinds of equipment that digitization projects typically require. This appendix provides a level of specificity which may not be essential to many project managers, but will indicate the level of detail needed to make certain kinds of purchasing and digitization decisions. Project managers need to understand the kinds of criteria that will be used for this kind of planning, even if they do not need to internalize the specifics.

Image and Video Equipment

Selecting the correct computer for digitization work is very important, but frequently overlooked. In particular image, video and audio processing is very processor intensive, as Ihrig and Ihrig point out, 'you never can have a fast enough computer or too much memory for scanning and processing images'. Therefore, at least one machine should be specified appropriately for this work. The key components to consider are memory, processor speed, the rate of data transfer between components, the size of the disk storage device, and audio and video cards.

When you plan to purchase a system you should test it to ensure that it meets your requirements. RAM-intensive operations such as image processing and video editing can take many seconds or even minutes in computers with dozens, rather than hundreds or even thousands of MB of memory. Design a test sequence to allow you to determine whether the machine will achieve your objectives.

Display devices

Do not neglect display where image work is concerned. Your computer display is likely to be the most used output device and a good quality display of sufficient size is important for effective and efficient quality control and comfortable working conditions. The display should ideally be 19" or 21" inch diagonal viewable measurement. Upgrading to a 21" display is a worthwhile investment. If the monitor is to be used for rigorous quality control ("soft proofing"), ensure that the display has hardware and/or software calibration, that the calibration mechanisms are compatible with the capture devices, and that standard color management profiles can be used. It is also important to control or calibrate the ambient environment along with the display device. If you are considering LCD flat panel monitors, they are available with analog or digital signal processors. The digital signal displays are better but not always compatible with all systems and video cards, so research the hardware carefully.

For many digital imaging projects, the resolution of the monitor can be far more important than the size of the screen. The monitor or screen resolution is the density of phosphor dots on the Cathode Ray Tube (CRT) or pixels in the matrix of a LCD monitor. This is usually measured in dot-pitch—e.g. 0.28mm dot-pitch is approximately 90 dots per inch (dpi).

Calculating Screen Resolution from Dot-Pitch:

You may find it useful to use dot-pitch numbers provided by computer suppliers to work out display dpi.

Screen Resolution = (1/dots per mm) x mm per inch
so a screen claiming a dot-pitch of 0.28 mm would have a resolution of 90dpi
90 = (1/0.28) x 25.4

Explanation--
0.28 mm 'pitch' means 1/0.28 dots per mm
1/0.28 dots per mm is equivalent to 3.5714 dots/mm
(constant 25.4 mm in an inch)
3.5714 * 25.4 =approx. 90 dots/inch

Relationship between Image dpi, Screen dpi and Magnification*:*
Level of Magnification = image dpi/screen dpi. Thus, regardless of the physical screen dimension (15" or 21" monitor), a 300 dpi image displayed on a 75 dpi monitor will be magnified 4:1.

Note that screen resolution can be measured in pixel dimensions, as well as dpi.

Pixel dimensions (both of image and monitor) are important if you are concerned about scrolling. (For example, if a monitor is set to a 1024 x 768 setting, regardless of the dpi of the image, if the image has more than 1024 pixels in its width or more than 768 pixels in its height, one will need to scroll in one or both directions.)

Relationship between Pixel Dimension Settings (controlled by software, not fixed by hardware) and "screen real estate":

Setting	Screen Real Estate	Gain
800 x 600	480,000 pixels	
1024 x 768	786,432 pixels	+63.8% over 800 x 600
1280 x 1024	1,310,720 pixels	+66.7% over 1024 x 768; +273% over 800 x 600
1600 x 1200	1,920,000 pixels	+46.5% over 1280 x 1024; +244% over 1024 x 768; +400% over 800 x 600

The lower the dot-pitch is, the higher the resolution of the monitor. Although the quality of flat panel LCD displays is improving and they are becoming increasingly affordable, in 2002 they do not yet achieve the same price-to-quality ratio of conventional CRT monitors. A final consideration in relation to display is the video adapter, or video card; often now built into the motherboard of computers you should expect to see at least 16MB of video RAM or more, as less than 16MB RAM will cause problems when viewing and manipulating large images.

Printers

When buying a printer, projects will need to consider the dimensions of the original material, the desired dimensions of the output, and whether color is required, as well as the printer resolution. Printer resolution is a measure of the printer's ability to produce individual dots, and is usually expressed as dots per inch (dpi). Printer resolution may also be expressed as lines per inch (lpi).

Color ink jet printers and associated media (inks, papers) have become very affordable and quality is often fit-for purpose. The variables of paper, ink, storage environment, and use all determine the life expectancy of their output. If one of the planned outcomes for the digitization project, however, is the ability to create the equivalent of photo prints directly from the digital source, it is best to consult the photography department at another institution or a reputable printing service bureau or advice on printer selection.

Digital imaging devices

Definition Box:

Capture Device	Details	Scanning Array
Digital Camera	Linear Sensor (tends to be slow)	Linear CCD Array
Digital Camera	1 shot cameras with color diodes	CCD Area Array
	3 shot cameras with color filters	CCD Area Array
Flatbed Scanner	Medium to high-end devices, better optical image quality	Linear CCD Array
	Low-end devices, poorer optical image quality. Rely on interpolation (software algorithms) to achieve their resolution	Linear CIS/CCD Sensor
Engineering Scanner	Media pulled past the scanning array rather than the array moving over the media. Typically range from 36" to 62" width capacity by up to _" think capacity.	Linear CCD Array
Drum Scanner	Typically high-end prepress or graphic house equipment. Needs specialist operator. Media tightly taped to a drum with a layer of oil and rotated round the PMT at high speed. Color correction of negatives can be problematic.	PMT
Hybrid Drum Scanner	Halfway between a drum scanner and a flatbed. Can handle a wider variety of media than transparency scanners. Media is held tight over a cylinder by a set of rollers. This has the advantages of a drum scanner, but without tape and oil.	CCD
Transparency Scanner	High resolutions suitable for 35mm negatives and slides. Higher end machines have more advanced software for accurate color correction of negatives.	CCD Area Array
Microfilm Scanner	Available for 16mm, 35mm and microfiche formats. Usually output in bi-tonal or grayscale. "Blipping", or the ability to distinguish individual frames, is the key to high throughput and accurate results.	Linear CCD Array
Book Scanner	Can be overhead (planetary) non-contact scanners or face down (contact) cradle scanners. Also suitable for large documents.	Linear CCD Array
Card Scanner	Designed for business card, check or fingerprint card scanning. Can be suitable for card catalogs.	Linear CCD Array

Audio Equipment

Institutions may find themselves with a rich array of materials in analog form, but without the devices to play this material back. Unlike textual and still image material (with the exception of slides and born digital), audio and moving image material require a playback device in addition to a digital capture device. For example, a flatbed scanner can digitize directly a wide range of reflective media of different formats and sizes (e.g., photographs, letters, printed matter, bus tickets). No similar general-purpose capture device for audio and moving image material exists. A collection that included 78 rpm records, compact cassettes, 8mm film and VHS video cassettes would require a playback device for each of these and each would then need to be connected to an appropriate digital capture device. For audio and moving image material that is already in a digital format (such as CD or Digibeta), playback equipment is less of a problem. Although many—frequently incompatible—proprietary digital formats exist, their recent development means suitable playback equipment is still on the market and relatively easy to source. Therefore this section concentrates on identifying analog audio and moving image formats, their properties and the source device required.

Three methods can be used to progress from moving image film to digital. Film can be transferred onto videotape for digitization via a transfer box or multiplexer. Both of these options depend upon the material being projected in some way. Transfer boxes project the image into a box containing a mirror and onto a rear image projection screen with the video camera mounted on the other side of the box. The resulting video is subsequently digitized. These transfer boxes are not expensive, but do not in general produce quality as high as a multiplexer because they introduce generational loss.

A better solution is to use a multiplexer. In this device the projector and camera are mounted on a single table. The image is projected by a set of lens and mirrors, directly into the camera without the need for a projection screen. This has advantages for image clarity. In both processes quality suffers because it introduces an extra production generation into the reformatting of the analog material. An alternative to these methods is the use of 8, 16 and 35mm film for a chain film scanner to digitize directly from the analog film material. These machines scan the films and digitize at the scanner, passing the digital signal to the computer. (They work slightly differently for digital video. In this instance they grab individual lines of video to construct a frame and produce broadcast quality digital video.) In 2001 the costs of these machines remains high at between $500,000 and $1,000,000. One of the strengths of chain scanning is that, because the analog to digital conversion is done at the camera rather than on the computer, there is less opportunity for noise to be added by the process to the analog signal. Whereas small institutions can probably set up a transfer box or multiplexer system, even wealthy institutions would find outsourcing to a facilities house to be the only practical option if they wished to go directly from the analog film to the digital material.

Determining a film's original frame rate is also difficult without viewing the film with a projector, particularly for old 8 and 16mm films. The widespread availability of VHS and S-VHS video players makes the playback of these video formats for digitization

relatively simple. The rapid adoption of digital formats in broadcasting, post-production and amateur markets is making the availability of even quite recent analog video devices scarce.

As there are fewer analog audio formats, these provide less of a problem than moving images. Compact cassette players, 33 and 45 rpm record players are still widely available new. Even record players with a 78 rpm speed can still be purchased new. The other formats present a greater challenge. If digitizing the sound as played on period equipment is important the tone arms of phonographs and gramophones can be customized to provide an appropriate feed. Alternatively, the sound can be recorded via an external microphone onto a more convenient intermediary format. Reel to reel tape, wire recorders and cartridges pose similar problems of transfer. By modifying the equipment, it may be possible to provide a direct sound output. Alternatively, the sound can again be captured via an external microphone to an appropriate intermediate format. Here is where a great deal of specialist advice can be helpful. Just as we noted that it is easier to train a good photographer in digitization than it is to train a digital expert in photographic principles and practice, you will find that sound engineers bring to the digital environment strengths that are difficult to replicate.

In the case of all audio and moving image material, whether it is in analog or digital form, projects should carefully consider the advantages of outsourcing digitization. In general audio and moving image digitization require more and more expensive and specialized equipment than is necessary for still image material.

Audio Media	Properties	Source Device
Wax or Celluloid Cylinders	1890s & 1900s, up to 5”diameter, 2-4 mins. playing time	Phonograph. See http://www.tinfoil.com for details of digital transfer.
Wire	Magnetic coated wire drums or reels. Invented 1898. Widely used by the US military in WWII. Eclipsed by magnetic tape by the mid 1950s.	Wire Recorder
78 rpm shellac resin discs	1898 to late 1950s, 10” (25cm) and 12” (30cm) most common sizes	Gramophone (wind-up) or Hi-Fi. Gramophone’s steel needles need replacing after each side or record played. Hi-Fi needs a 78 rpm turntable and a cartridge with a 78 rpm stylus. For best results on modern equipment a phono pre-amplifier is required to correctly equalize the different types of record.
45 rpm and 33 rpm vinyl discs	7” (20cm) single and 12” (30cm) long play. Long play (LPs) introduced in 1948, stereo recordings in 1958.	Hi-Fi. Hi-Fi requires turntable with 45 and 33 rpm speeds.
Reel to Reel magnetic tape	1/2” to 1/4” magnetic tape. BASF and AEG developed 6.5mm ferric tape and Magnetophone player in Germany from 1935. Post-war development in USA by Ampex and 3M. Stereo capability from 1949.	Reel to Reel player for appropriate width of tape.
Compact Cassette	Magnetic polyester tape introduced by Philips in 1963.	Hi-Fi. Hi-Fi requires compact cassette player.
Cartridge	1/4” magnetic tape. Fidelipac (4-track, devised 1956, released 1962) and Lear (8-track, 1965) cartridge systems.	Despite similarities 4 and 8 track cartridges are not compatible and require separate players. Predominantly used for in-car audio. 4 track unpopular outside of California and Florida.

Storage devices and systems

File storage requirements are a major consideration with any digitization project, and particularly for projects dealing with digital images or video. While storage devices have grown steadily since their invention and show no signs of reaching a plateau, storage demands have kept pace with them, and file storage remains a concern for delivery and for backup. Consider that 150 24-bit color TIFF files at 400 pixels per inch (ppi), plus 150 JPEG files at 100 ppi, together with associated metadata, would occupy 5 - 6 GB. In 2002, this is approximately the entire capacity of most standard PC internal hard drives. Different types of analog material (text, images, audio) and different ways of encoding it will have a direct impact on the sizes of the files and the quantities of storage that a project will require. A terabyte, or 1000 gigabytes, offers sufficient storage capacity to hold the equivalent of 300,000,000 pages of ASCII text, 20,000,000 pages of bitonal scanned documents, 1,000,000 pages of color images, 1,800 hours of recorded sound, or 500 hours of good quality video.

Projects have various storage options, of which the most common are internal hard drives, optical drives, tape drives and networked hard drives. Internal computer hard

drives are fairly fast and capacious, but since they receive daily use they are at risk for hardware failure, infection by viruses, and similar calamities. Since they do not use a removable medium, they need to be backed up onto some other storage format regularly. Optical drives (including CD-R and DVD-R) are removable and fairly durable, although their capacity is limited (640 Mb for CD-R, 6 Gb for DVD-R). Tape drives provide much greater capacity, but their access speed is comparatively slow and the tapes require care and protection to avoid loss of data. There are at least two types of digital tape: DAT (digital audio tape) is the slower and more unreliable of the two, and offers a somewhat lower capacity (on the order of 12-20 Gb without compression), but uses cheaper equipment. DLT (digital linear tape) is faster, more reliable, and offers capacity in the range of 35-70 Gb without compression, but the equipment is more expensive. A newer format, AIT (Advanced Intelligent Tape) offers even greater capacity but slow speeds. Networked hard drives offer increased speed and capacity; in the case of a RAID array (which uses multiple drives acting in tandem) there is also the possibility of complete data redundancy and protection against loss. These are also among the most expensive options when expressed as cost per megabyte of storage.

Another distinction worth examining is the difference between online and offline storage devices, each of which has its own advantages and special considerations. Online storage (i.e. storage which, like a disk drive, is always attached to a live computer), provides immediate “random access” to any part of the storage system. Offline storage (or removable storage, such as tapes or CD-R disks) typically provides a lower cost per Mb of storage, but has the disadvantage that when you want to access a file you must locate the media on which it is stored, fetch the media, and load it. It is also typically slower in actual access speed (the speed with which data can be read from the disk or tape) than online storage. Hierarchical storage systems (or Storage Management Systems—SMS) combine elements of online and offline storage along with special control software to provide seamless access to a combined system.

Each of these options is ideal for some purposes and not for others, so when purchasing storage you should consider where your needs chiefly lie. For backup and preservation purposes, it is essential that the storage medium be removable and fairly durable; depending on how much data you need to back up, it may also need to be high-capacity. For smaller projects, or projects dealing with smaller files (such as text files in SGML/XML), backing up onto CD-ROM or DVD may be sufficient. (Remember that you should regularly back up not only your actual digital assets, but also all of your work files, including administrative materials, tracking records, correspondence, and so forth; be sure to include these in your estimates of required backup capacity.) For high-capacity backup, such as may be required by large image or video projects, tape storage may be necessary. The requirements for preservation are very similar, with the proviso that reliability and long-term durability may be even more important factors to consider. Most of the projects surveyed for this *Guide* use either CDs or tape for external backup.

For publication purposes (i.e. for storing the data that is being actively delivered online) speed is essential; no removable storage devices are sufficiently fast. For large projects with substantial amounts of data and high demand, a high-speed, high-capacity system

such as a RAID will be required. A RAID (Redundant Array of Independent Disks) is a collection of drives acting in tandem, which offers not only very high read-write speed, but also the option of full data redundancy (if one drive fails, the system can “fail over” to another drive seamlessly, and the failed drive can be replaced without interrupting data access). Such systems are very expensive, but if your data is mission-critical and you cannot afford any down time, their reliability is worth the investment. It may also be worth considering outsourcing storage of this type, or working with a consortium to share storage costs among several institutions.

Projects may also wish to consider data warehousing, a data repository specifically structured for querying and reporting. Data warehousing is a process used by large businesses to ensure sustainability and access to data about the company but cultural and heritage institutions may find the process or even just the concept, to be useful in storage planning. Several variables are worth mentioning here:

- Subject-oriented - Data that gives information about a particular subject instead of about a company’s ongoing operations.
- Integrated - Data that is gathered into the data warehouse from a variety of sources and merged into a coherent whole.
- Time-variant - All data in the warehouse is identified with a particular time period.
- Non-volatile - Data is stable in a data warehouse. More data is added, but data is never removed. This enables management to gain a consistent picture of the business.

Appendix B: Metadata

This appendix brings together material from various sections of the *Guide*, in expanded form, to provide a detailed description of the kinds of metadata and metadata standards that are of greatest importance to the cultural heritage sector. It has already been observed, but is worth repeating, that metadata is a crucial part of any cultural heritage digitization project and should not be neglected. Without metadata identifying its source, contents, and details of creation at an absolute minimum, a digital object is useless. Capturing additional information to facilitate rights management, administrative tracking, preservation, and distribution can enable you to get much more powerful use from your digital materials with much less difficulty.

What is metadata?

Metadata is literally "information about the data": information created about the source material and the digital version to record the essentials of their identity, creation, use, and structure. Their purpose is to facilitate the discovery, use, management and reusability of digital material. Metadata can be usefully divided into three categories: descriptive, administrative and structural. These are not rigidly bounded groups and they frequently overlap.

Descriptive metadata describes and identifies information resources, to facilitate searching, retrieval, and management. It typically includes basic bibliographic information such as the creator, title, creation date; catalog information such as accession or other identifying numbers; and topic information such as keywording. Examples of descriptive metadata include Library of Congress Subject Headings, Categories for the Description of Works of Art (CDWA) (http://www.getty.edu/research/institute/standards/cdwa/), the Art and Architecture Thesaurus (AAT) (http://www.getty.edu/research/tools/vocabulary/aat/), and the Dublin Core metadata set.

Administrative metadata is used to facilitate management, tracking, migration and re-use of digital assets. It typically includes information on creation, quality control, rights and preservation. See Cornell University's "Web Hub for Developing Administrative Metadata for Electronic Resource Management" (http://www.library.cornell.edu/cts/elicensestudy/home.html). The term "technical metadata" is also used in a similar sense to indicate metadata about data capture and the technical characteristics of the images.

Structural metadata describes the internal structure of digital resources and the relationships between their parts. It is used to enable navigation and presentation. Examples of structural metadata are included in the METS standard

(http://www.loc.gov/standards/mets/) and SMIL (Synchronized Multimedia Integration Language) (http://www.w3.org/TR/REC-smil/)

Text metadata

Metadata for textual resources is in some ways the most straightforward to create, because it can be captured in the same format as the digital object itself, and can be included directly in the digital file, for instance as a header section in an SGML/XML-encoded document. EAD, TEI, and HTML all include a header element of varying scope; the HTML header provides for the inclusion of basic Dublin Core metadata terms, while the EAD and TEI headers provide for more extensive information about both the electronic file and the information it captures or describes.

Image metadata

Metadata for still images may be stored in a file header or in a separate database or file. Because images themselves, unlike text, cannot currently be searched very effectively for content, metadata is doubly important for retrieval purposes, as well as for internal project tracking, management, and documentation. The METS standard can be used to bring together the different types of image metadata required for different project purposes. It can also be used not only to document individual images, but also to represent the relationships between multiple images that together constitute a single digital object (for instance, high-resolution archival images, thumbnails and delivery images at lower resolutions, images of particular details at higher magnification). The NISO IMG draft standard on metadata requirements for digital still images provides extremely detailed specifications for capturing technical metadata for images: http://www.niso.org/standards/dsftu.html and http://www.niso.org/standards/resources/Z39_87_trial_use.pdf

Audio-visual metadata

As with still images, metadata is crucial to digital audio or video, and the principles of metadata interoperability and documentation standards are as important to digital AV media as to still image and text media. Metadata for digital audio and visual resources can be used in much the same way as metadata for complex digital objects composed of still images. A metadata standard like METS (with the appropriate extension schema) can be used to describe the structure of an audio-visual digital object: for instance, a group of original taped interviews and a final version edited for delivery.

SMIL (Synchronized Multimedia Integration Language) can be used to describe the content and structure of time-based digital files such as audio and video. SMIL, for instance, can be used to describe structural metadata about a particular frame of video (frame 30, timecode 01:20:36.01) as well as to link the appropriate series of frames to

alternate representations such as a transcription of the dialogue in that scene. As with image resources, this allows users to search for a particular bit of dialogue or the name of a character, and be taken directly to the video scene in which they appear.

Link Box:

Key AV Metadata Sites

- Dublin Core Metadata Implementers: http://www.fiu.edu/~diglib/DC/impPurpose.html
- Synchronized Multimedia Integration Language (SMIL): http://www.w3.org/AudioVideo/
- Metadata Encoding and Transmission (METS) Standard: http://www.loc.gov/mets
- MPEG-7: http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm
- Authority Tools for Audio-Visual Catalogers: http://ublib.buffalo.edu/libraries/units/cts/olac/capc/authtools.html#g
- Authority Resources for Cataloging Popular Music: http://www.music.indiana.edu/tech_s/mla/wgpms/wgpms.htm
- Library of Congress's Digital Audio-Visual Preservation Prototyping Project: http://lcweb.loc.gov/rr/mopic/avprot/avlcdocs.html#md
- Library of Congress's Digital Audio-Visual Extensions to METS Standard: http://www.loc.gov/rr/mopic/avprot/metsmenu2.html
- Cinemedia's SWIFT project for on-demand delivery of film and video: http://www.cinemedia.net/SWIFT/project.html

Metadata standards

A number of metadata standards are now in use by the cultural heritage community that have been developed by different subcommunities to address particular needs. These standards are not mutually exclusive; on the contrary, some of them, such as METS, are specifically intended to be a way of bringing together various forms of metadata in a single place where it can be processed uniformly and predictably.

METS

The Metadata Encoding and Transmission Standard (METS) is an XML-based encoding standard for digital library metadata. It is both powerful and inclusive, and makes provision for encoding structural, descriptive, and administrative metadata. It is designed not to supersede existing metadata systems such as Dublin Core or the TEI Header, but rather to provide a way of referencing them and including them in the METS document. As a result, it is an extremely versatile way of bringing together a wide range of metadata about a given digital object. Through its structural metadata section, it allows you to express the relationships between multiple representations of the digital object (for instance, encoded TEI files, scanned page images, and audio recordings), as well as relationships between multiple parts of a single digital representation (for instance, the

sections of an encoded book). Its administrative metadata section supports the encoding of the kinds of information projects require to manage and track digital objects and their delivery: technical information such as file format and creation; rights metadata such as copyright and licensing information; information about the analog source; and information on the provenance and revision history of the digital objects, including any data migration or transformations which have been performed. METS is a very recently developed standard but is well worth watching and using.

Dublin Core

The Dublin Core Metadata Element Set defines a set of 15 essential metadata components (for instance, author, title, format) which are broadly useful across disciplines and projects for resource discovery and retrieval. These components can be used to add metadata to HTML files (using the <meta> tag) but can also be used in other contexts to create basic metadata for a wide range of digital resources. Dublin Core does not provide for detailed administrative or technical metadata, and as such is largely suited for exposing resources for search and retrieval, rather than for internal resource management and tracking. In addition, since its goal is to be simple and broadly applicable to a wide variety of resources, it does not provide for the kind of highly structured metadata about specific document types that TEI and EAD offer. Although projects using these encoding systems will probably not need to use the Dublin Core, they may find it useful to be aware of it as a possible output format for distributing metadata about their resources. One aspect of the work of the Consortium for the Interchange of Museum Information (CIMI) is research into SGML, XML, and metadata standards such as Dublin Core for museum collections and RDF.

TEI Header

The TEI Header is a required component of any file conforming to the Text Encoding Initiative Guidelines, and is ordinarily used to document a text file encoded in TEI. However, it can also be used to describe other kinds of resources. It is designed to express a wide range of metadata about a digital file, whether that file is an encoded text, an image, a digital recording, or a group of any of these. It provides not only for standard bibliographic information about the file itself and about its source, but also more specialized metadata to record the details of classification schemes, encoding and sampling systems used, linguistic details, editorial methods, and administrative metadata such as the revision history of the file. It is designed to accommodate a wide range of metadata practices, and while it offers highly structured options for capturing detailed metadata, it also allows for briefer and more loosely organized headers which record only the most basic information.

EAD

In a sense, the Encoded Archival Description (EAD) bridges the realms of data and metadata. As a digital finding aid, it may stand on its own as metadata about an archival collection. As a digital representation of an analog finding aid, it may also be a form of

digital preservation (particularly if the original finding aid has any historical significance). It provides for the capture of all the information ordinarily conveyed in a finding aid, but it also provides for metadata about the finding aid itself-- its author, language, publication details-- and about the EAD file as well. EAD is a powerful tool for providing digital access to archival collections by representing the information user's need to discover archival materials of interest in a consistent and digitally transparent way.

SPECTRUM

The UK Museum Documentation Standard represents a common understanding of good practice for museum documentation, established in partnership with the museum community. It contains procedures for documenting objects and the processes they undergo, as well as identifying and describing the information which needs to be recorded to support the procedures. Spectrum was developed by the MDA in Great Britain. CIMI has adapted the SPECTRUM XML DTD for web based museum object.

Appendix C: Digital Data Capture

This appendix brings together material from various sections of the *Guide*, in expanded form, to provide a detailed description of how analog information is converted into digital data in various media types. While for many purposes this level of technical detail may be more than is needed, a basic understanding of the principles involved can be useful in evaluating the appropriateness of certain types of equipment, determining when digitization is likely to yield good results (or not), and understanding why certain kinds of conversion can result in data loss or degradation. Specific recommendations for formats, settings, and how to get the best results from different kinds of materials are addressed in the main sections of the *Guide*; the goal here is to provide a more detailed explanation of the basic principles involved.

General Principles

Analog and digital data are fundamentally different: where analog information is generally smooth and continuous, digital information consists of discrete chunks, and where analog information bears a direct and non-arbitrary relationship to what it represents, digital information is captured using formal codes that have only an arbitrary and indirect relationship to the source. Thus while an analog image, for instance, consists of continuously varying colors and shading, a digital image consists of a set of individual dots or pixels, each recording the color intensity and other information at a given point. Although the specific kinds of information vary from medium to medium—sound waves, light intensities, colors—this basic difference remains a constant.

Conversion from analog to digital thus requires that the continuous analog information be sampled and measured, and then recorded in digital format. There are several basic factors which govern this process and which determine the quality of the digital result.

The first of these is the density of data being captured from the analog original: in effect, how often the original is sampled per unit of time (in the case of video and audio) or area (in the case of images and video). For digital audio, the higher the sampling rate, the smoother the transitions between the individual packets of sound, to the point where, with modern digital audio, they cannot be detected by the human ear. A low sampling rate results in clipping, the audio equivalent of jerky animation. For digital images, the higher the sampling rate (i.e. resolution), the smoother and less pixellated the image appears, and the more it can be magnified before its granularity becomes visible.

The second factor at work is the amount of information that is recorded in each sample. Individual pixels in an image may contain very little information-- at the most minimal, they may take only one binary digit to express on versus off, black and white—or they may take 32 bits to express millions of possible colors. Large sample size may be used, as

in digital images, to capture nuance, finer shadings of difference between values. They may also be used to express a wider total range, as in the case of digital audio, where a higher frequency response means that the recording can capture a greater range of frequencies, with higher highs and lower lows.

Both sampling frequency (or resolution) and sample size (frequency response, bit-depth) involve a trade-off of data quality and file size. It is clear that the more frequently you sample, and the more information you capture in each sample, the larger your file size will be, and the more costly to create, transmit, store, and preserve. Decisions about digital data capture are thus not simply a matter of achieving the highest possible quality, but rather of determining the quality level that will represent the original adequately, given your needs. Various sections of the *Guide* explore these considerations in more depth.

The remainder of this appendix describes how these principles apply in detail in particular digital media.

Digital Audio and Video Capture

In analog audio recording, a plucked string (for example) vibrates the air around it. These airwaves in turn vibrate a small membrane in a microphone and the membrane translates those vibrations into fluctuating electronic voltages. During recording to tape, these voltages charge magnetic particles on the tape, which when played back will duplicate the original voltages, and hence the original sound. Recording moving images works similarly, except that instead of air vibrating a membrane, fluctuating light strikes an electronic receptor that changes those fluctuations into voltages.

Sound pressure waveforms and other analog signals vary continuously; they change from instant to instant, and as they change between two values, they go through all the values in between. Analog recordings represent real world sounds and images that have been translated into continually changing electronic voltages. Digital recording converts the analog wave into a stream of numbers and records the numbers instead of the wave. The conversion to digital is achieved using a device called an analog-to-digital converter (ADC). To play back the music, the stream of numbers is converted back to an analog wave by a digital-to-analog converter (DAC). The result is a recording with very high fidelity (very high similarity between the original signal and the reproduced signal) and perfect reproduction (the recording sounds the same every single time you play it no matter how many times you play it).

When a sound wave is sampled using an analog-to-digital converter, two variables must be controlled. The first is the sampling rate, which controls how many samples of sound are taken per second. The second is the sampling precision, which controls how many different gradations (quantization levels) are possible when taking the sample. The fidelity of the reproduced wave can never be as accurate as the analog original; the difference between the analog signal and the closet sample value is known as quantization error. This error is reduced by increasing both the sampling rate and the

sampling precision. As the sampling rate and quantization levels increase, so does perceived sound quality.

In digital representation, the same varying voltages are sampled or measured at a specific rate, (e.g. 48,000 times a second or 48 kHz). The sample value is a number equal to the signal amplitude at the sampling instant. The frequency response of the digital audio file is slightly less than half the sampling rate (Nyquist Theorem). Because of sampling, a digital signal is segmented into steps that define the overall frequency response of the signal. A signal sampled at 48 kHz has a wider frequency response than one sampled at 44.1 kHz. These samples are represented by bits (0's and 1's) that can be processed and recorded. The more bits a sample contains, the better the picture or sound quality (e.g., 10-bit is better than 8-bit). A good digital signal will have a high number of samples (e.g., a high sampling rate) and a high number of bits (quantizing). Digital to digital processing is lossless and produces perfect copies or clones, because the digital information can be copied with complete exactness, unlike analog voltages. High bit-depth is also result in much-increased dynamic range and lower quantization noise.

Ideally, each sampled amplitude value must exactly equal the true signal amplitude at the sampling instant. ADCs do not achieve this level of perfection. Normally, a fixed number of bits (binary digits) is used to represent a sample value. Therefore, the infinite set of values possible in the analog signal is not available for the samples. In fact, if there are R bits in each sample, exactly 2R sample values are possible. For high-fidelity applications, such as archival copies of analog recordings, 24 bits per sample, or a so-called 24 bit resolution, should be used. The difference between the analog signal and the closest sample value is known as quantization error. Since it can be regarded as noise added to an otherwise perfect sample value, it is also often called quantization noise. 24-bit digital audio has negligible amounts of quantization noise.

Digital Image Capture

Digital image capture divides the image into a grid of tiny regions, each of which is represented by a digital value which records color information. The resolution of the image indicates how densely packed these regions are and is the most familiar measure of image quality. However, in addition to resolution you need to consider the bit-depth, the amount of information recorded for each region and hence the possible range of tonal values. Scanners record tonal values in digital images in one of three general ways: black and white, grayscale, and color. In black and white image capture, each pixel in the digital image is represented as either black or white (on or off). In 8-bit grayscale capture, where each sample is expressed using 8 bits of information (for 256 possible values) the tonal values in the original are recorded with a much larger palette that includes not only black and white, but also 254 intermediate shades of gray. In 24-bit color scanning, the tonal values in the original are reproduced from combinations of red, green, and blue (RGB) with palettes representing up to 16.7 million colors.

Digital Text Capture

Although it may seem odd to discuss digital text in this context, there are some important, if indirect parallels between the principles described above and those that govern digital text capture. Clearly in capturing digital text one does not sample the original in the same way that one samples audio or images. However, the process of text capture does involve choices about the level of granularity at which the digital representation will operate. In capturing a 20th-century printed text, for instance, a range of different "data densities" is possible: a simple transcription of the actual letters and spaces printed on the page; a higher-order transcription which also represents the nature of textual units such as paragraphs and headings; an even more dense transcription which also adds inferential information such as keywords or metrical data. Other possibilities arise in texts that have different kinds of internal granularity. In the case of a medieval manuscript, one might create a transcription that captures the graphemes—the individual characters—of the text but does not distinguish between different forms of the same letter (for instance, short and long s). Or one might capture these different letter forms, or even distinguish between swashed and unswashed characters. One might also choose to capture variations in spacing between letters, lines of text, and text components, or variations in letter size, or changes in handwriting, or any one of a number of possibly meaningful distinctions.

These distinctions, and the choice of whether or not to capture them, are the equivalent of sampling rates and bit-depth: they govern the amount of information which the digital file records about the analog source, and the resulting amount of nuance that is possible in reusing and processing the digital file.

References

American Memory Evaluation Team. “Final Report of the American Memory User Evaluation, 1991-1993.” 1993, American Memory Project, Library of Congress, http://memory.loc.gov/ammem/usereval.html, (acc. September 2002).

Bearman, D., G. Rust, S. Weibel, E. Miller, and J. Trant. “A Common Model to Support Interoperable Metadata. Progress Report on Reconciling Metadata Requirements from the Dublin Core and INDECS/DOI Communities.” *D-Lib Magazine* 5, no. 1, January (1999). http://www.dlib.org/dlib/january99/bearman/01bearman.html, (acc. Oct 2000).

Emery, P. "The Content Management Market: What You Really Need to Know." *Spectra*, vol. 29, no 1 (2002), pp.34-38.

Fraser, B. F. Bunting, and C. Murphy. *Real World Color Management*. Peachpit Press: forthcoming, autumn 2002.

Friedlander, A. “The National Digital Information Infrastructure Preservation Program: Expectations, Realities, Choices and Progress to Date,” *D-Lib Magazine* 8, no. 4, April (2002). http://www.dlib.org/dlib/april02/friedlander/04friedlander.html.

Gay, G., and R. Rieger. “Tools and Techniques in Evaluating Digital Imaging Projects”. *RLG Diginews* 3, no 3, June 15 (1999). http://www.rlg.org/preserv/diginews/diginews3-3.html#technical1 (acc.September 2002).

Hazen, D., J. Horrell, and J. Merrill-Oldham. *Selecting Research Collections for Digitization*. Council on Library and Information Resources (August 1998). http://www.clir.org/pubs/reports/hazen/pub74.html.

Johnson, N. F., and S. Jajodia. “Exploring Steganography: Seeing the Unseen.” *Computer* February 31, no. 2 (1998): 26-34.

Kiernan, K. “Digital Preservation, restoration, and the dissemination of medieval manuscripts.” *Gateways, gatekeepers, and roles in the information omniverse: proceedings of the third symposium: November 13-15, 1993*, edited by A Okerson and D Mogge, Washington, DC, 1994.

Lesk, M. *Image Formats for Preservation and Access: A Report of the Technology Assessment Advisory Committee to the Commission on Preservation and Access*. Washington, D.C.: Commission on Preservation and Access, 1990.

Lesk, M. Report on *“Real World”* searching panel at SIGIR97. ACM SIGIR Forum, 32(1):1-4, Spring 1998.

Mintzer, F. C., L. E. Boyle, A. N. Cazes, B. S. Christian, S. C. Cox, F. P. Giordano, H. M. Gladney, J. C. Lee, M. L. Kelmanson, A. C. Lirani, K. A. Magerlein, A. M. B. Pavani, and F. Schiattarella. "Toward On-line, Worldwide Access to Vatican Library Materials." *IBM Journal of Research and Development* 40, no. 2 (1996). http://www.research.ibm.com/journal/rd/mintz/mintzer.html, (acc. Oct 2000).

Mudawwar, M. F. "Multicode: A Truly Multilingual Approach to Text Encoding." IEEE Computer 30(4): 37-43 (1997).

OCLC/RLG. "Preservation Metadata for Digital Objects: A Review of the State of the Art." A white paper by the OCLC/RLG Working Group on Preservation Metadata, January 2001, http://www.oclc.org/research/pmwg/presmeta_wp.pdf.

Prescott, A. "Constructing Electronic Beowulf." *Towards the Digital Library: The British Library's 'Initiatives for Access' Programme*, edited by Leona Carpenter, Simon Shaw and Andrew Prescott. London: The British Library, 1998. 30-49.

Puglia, S. "The Costs of Digital Imaging Projects." *RLG DigiNews* 3, no. 5, October 15 (1999). http://www.rlg.org/preserv/diginews/diginews3-5.html#feature, (acc. Oct 2000).

Ross, S. "Strategies for Selecting Resources for Digitization: Source-Orientated, User-Driven, Asset-Aware Model (SOUDAAM)." *Making Information Available in Digital Format: Perspectives from Practitioners*, edited by Terry Coppock, Edinburgh: The Stationary Office, 1999. 5-27.

Ross, S. and Economou, M. 1998. 'Information and Communications Technology in the Cultural Sector: The Need for National Strategies', *DLib Magazine* June 1998. http://www.dlib.org/dlib/june98/06ross.html.

Royan, B. "Cross-domain access to digitised cultural resources: the SCRAN project." 64th IFLA General Conference, Amsterdam, 16-21 August 1998. http://ifla.inist.fr/IV/ifla64/039-109e.htm (acc. 8 March 2002).

Royan, B. "Scotland in Europe: SCRAN as a Maquette for the European Cultural Heritage Network." *Cultivate Interactive* July 2000. http://www.cultivate-int.org/issue1/scran/ (acc. 2002)

Shapiro, M. and B. Miller. *A Museum Guide to Copyright and Trademark*. American Association of Museums, 1999. http://www.aam-us.org/resources/reference_library/mus_guide_copyright.cfm (excerpts).

Smith, A. *The Future of the Past: Preservation in American Research Libraries*. Washington, DC: Council on Library and Information Resources (CLIR), 1999. http://www.clir.org/pubs/reports/pub82/pub82text.html, (acc. Oct 2000).

Waibel, G. “Produce, Publish and Preserve: A Holistic Approach to Digital Assets Management.” Berkeley Art Museum Pacific Film Archive, n.d. http://www.bampfa.berkeley.edu/moac/imaging/index.html

Zorich, D. M. *Introduction to Managing Digital Assets: Options for Cultural and Educational Organizations*. Los Angeles: Getty Information Institute, 1999.

Zorich, D. “Why the Public Domain is Not Just a Mickey Mouse Issue,” NINCH Copyright Town Meeting, Chicago Historical Society, January 11, 2000. http://www.ninch.org/copyright/2000/chicagozorich.html

Abbreviations Used in the *Guide*

AAT	Art and Architecture Thesarus or Applications of Advanced Technologies Program
AHDS	Arts and Humanities Data Service
AIF	Audio Interchange File Format
AMICO	Art Museum Image Consortium
ASCII	American Standard Code for Information Interchange
BAMPFA	Berkeley Art Museum and Pacific Film Archive
BL	British Library
CCP	Classics Computer Project
CD	Compact Disc
CDL	Californian Digital Library
CDP	Colorado Digitization Project
CIDC	Cornell Institute for Digital Collections
CIMI	Consortium Museum Intelligence
CITI	Chicago Information, Tracking and Inquiry
CMS	Collection Management Services (Division, Library of Virginia)
CMYK	Cyan Magenta Yellow and Black
CNN	Cable News Network
CTIT	Centre for Telematics and Information Technology
DAT	Digital Audio Tape
DC	Dublin Core
DFG	Deutsche Forschungsgemeinschaft
DIMTI	Digital Imaging and Media Technology Initiative
DLIB	Digital Library
DLIT	Digital Library and Information Technologies
DLP	Digital Library Program
DLPS	Digital Library Production Service
DLXS	Digital Library eXtension Service
dpi	Dots per inch
DTD	Document-Type Definition
DVD	Digital Versatile Disc
EAD	Encoded Archival Description
EAF	Early American Fiction
FTE	Full Time Equivalent
FTP	File Transfer Protocol
G&M	Geography and Map (Division, Library of Congress)
GDZ	Göttinger DigitalisierungsZentrum
GIF	Graphics Interchange Format
GIS	Geographical Information Systems
GPS	Global Positioning System
GSU	Genealogical Society of Utah
HATII	Humanities Advanced Technology and Information Institute

HLF	Heritage Lottery Fund
HP	Hewlett-Packard
HTML	Hypertext Markup Language
IAIA	Integrated Arts Information Access project
IATH	Institute for Advanced Technology in the Humanities
IBM	International Business Machines
IMLS	Institute of Museum and Library Studies
IPR	Intellectual Property Rights
ITC	Information Technology and Communication
ITS	Information Technology Service
JPEG	Joint Photographic Experts Group
LAN	Local Area Network
LC	Library of Congress
LCSH	Library of Congress Subject Headings
LDI	Library Digital Initiative
LV	Library of Virginia
LZW	Lempel Ziv Welch
MAC	Apple MacIntosh Computer
MARC	Machine Readable Catalogue
MESL	Museum Educational Site Licensing
MIA	Minneapolis institute of Art
MIDI	Musical Instrument Digital Interface
MLS	Masters Library Science
MOA	Making of America
MOAC	Museums and the Online Archive of California
MOV/AVI	QuickTime Movie File Format
MPEG	Moving Picture Experts Group
MrSID	TIFF file viewer
MS	Microsoft
NDLP	National Digital Library Program
NEH	National Endowment for the Humanities
NMR	National Monuments Record (NMR
NRK	National Broadcasting Corporation (Norway)
NT	New Technology (Microsoft Operating System)
NTSC	National Television System Committee
NYPL	New York Public Library
OAC	Online Archive of California
OCR	Optical Character Recognition
OPAC	Online Public Access Catalogue
P&P	Prints and Photographs (Division, Library of Congress)
PAL	Phase Alternation Line
PC	Personal Computer
PDF	Portable Document Format
PEAK	Pricing Electronic Access to Knowledge
PM	Project Management
PT	Part Time

QC	Quality Control
QT	QuickTime
Ra	Real Audio/Video
RDF	Resource Description Framework
RGB	Red Green Blue
RSAMD	Royal Scottish Academy of Music and Drama
RTF	Rich Text Format
SCRAN	Scottish Cultural Resource Access Network
SGML	Standard Generalized Markup Language
SIBL	Science, Business & Industry Library (NYPL)
SLR	Single Lens Reflex
STG	Scholarly Technology Group
TEI	Text Encoding Initiative
TIFF	Tagged Image File Format
TLG	Thesaurus Linguae Graecae
TML	Thesaurus Musicarum Latinarum
TOC	Table of Contents
UK	United Kingdom
ULAN	Union List of Artists Names
UMICH	University of Michigan
UNESCO	United Nations Education Social Commission
US	United States
USD	United States Dollars
USMARC	United States Machine Readable Catalog
UV	Ultra Violet
UVA	Virginia University
VRA	Visual Resources Association
VHF	Visual History Foundation
VIA	Visual Information Access
VSL	Virginia State Library
VTLS	Virtual Library System
W3C	The World Wide Web Consortium
WAC	Walker Art Center
WAN	Wide Area Network
WAV	Wave format
XML	Extensible Markup Language