AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship

January 2012

University of Hull
Stanford University
University of Virginia
Yale University

Acknowledgement

The AIMS Project is a partnership between the University of Virginia Libraries, Stanford University Libraries and Academic Resources, the University of Hull Library, and Yale University Library with support from the Andrew W. Mellon Foundation.

Suggested citation:

AIMS Work Group. 2012. AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship. http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final_A4.pdf.

Table of Contents

Foreword

The AIMS project evolved around a common need among the project partners — and most libraries and archives — to identify a methodology or continuous framework for stewarding born-digital archival materials. These materials have been slowly accumulating in archival backlogs for years but are rapidly growing as more contemporary collections are accessioned.

Alongside the many and complex technological requirements, the challenges of stewarding born-digital material demand new strategies as well as a redefinition of archival workflows. Accordingly, this emerging challenge will affect the skill-set needed for archivists and the working relationships among archival colleagues as well as those outside our communities and organizations. If the archival profession aims to preserve and manage born-digital material to standards matching those of paper-based collections, a broader and deeper understanding of these issues must be developed, and this understanding must be incorporated into training of new archival professionals, professional development programs, and continuing education.

In both the United Kingdom and the United States — the home countries of the AIMS partners — there is a perception of a high bar for entry in the world of digital archives, both in terms of expertise and resources. Therefore, many institutions are reluctant to take even initial steps.

In the US in particular, organizational cultures have made sharing best practices difficult. While the electronic records, or e-records, community in the US has focused more on organizational records from information and knowledge management perspectives, those working in manuscript collecting repositories have been somewhat reluctant to enter an unfamiliar arena. Common issues in these collecting repositories — for example, legacy material and undefined accessioning practices — made it difficult to build expertise and capacity. Moreover, institutional practice has been focused on immediate local needs rather than developing a shared framework.

Now there is a small but emerging group of archivists working on issues related to born-digital content in personal papers and committed to sharing best practices. In addition, there is a growing recognition between archivists and those in the digital community that collaboration is absolutely crucial to success in this new paradigm.

The size of the archival community in the UK makes for a smaller arena within which to share ideas and solutions. In the UK there is a more developed, even thriving, community of practitioners working on born-digital archives of external donors/depositors as well as from their own organizations. However, there is a wide and growing gap between institutions with established staff, equipment, and processes (mostly national institutions and some universities) and those with no expertise or capacity whatsoever. Many smaller repositories cannot afford to collaborate with other institutions and thus cannot share some of the developments of their better-funded colleagues.

Despite these challenges, individual institutions and collaborative partnerships in both the UK and US are doing a great deal of work in research, development, and practical implementation. Some of the many projects and initiatives that influenced and informed the work of the AIMS partners are discussed in the next section. These projects approach the issues from different archival and technical perspectives. Recently, new tools have been developed that focused on capture, identification, or preservation. Some have discovered and are incorporating tools used in other fields, particularly technology developed for forensic investigations.

Although a great deal of work has been and continues to be done in this area, there is not yet a unified approach to address the lifecycle of stewardship in an accessible way — and most importantly, in a way that is grounded in archival practice. There is no single model to evaluate these many different approaches to born-digital stewardship and to unite them in a framework of objectives and options.

The AIMS Project

Into this climate, the AIMS partners proposed an inter-institutional framework for stewarding born-digital content. The AIMS partners realized that they could not solve all problems associated with born-digital materials but decided to focus their attention on professional practice defined by archival principles and by the current state of collections at the partner institutions.

In developing the AIMS Framework, the project would apply a practitioner-based research approach by developing a model based on real case studies of collections at each institution. Applying our theories would confirm or challenge the initial framework which could then be used as a model around which to build individual workflows and processes within each partner's organization. This test of concept for the AIMS Framework would prove whether it could be used within a wide range of organizations with different staffing models, archival processes, tools and infrastructure. This practical approach imposed a discipline and a framework for investigations and discussions; provided a variety of case studies, with different record formats, legacy issues, scale and complexity, and donor relationships; and defined an archival context for identifying ethical issues and other challenges, clearly demonstrating the need for workable and scaleable solutions.

The AIMS project was originally tasked to make recommendations for best practice including tools and workflows which could be applied within a variety of institutional scenarios. At a relatively early stage, however, it became clear that the development of best practice within born-digital stewardship was not yet possible. Tools do not yet exist for many elements of archival practice and many workflows are influenced by constantly changing institutional factors such as staff and technological infrastructure. The AIMS Framework, therefore, was developed to define good practice in terms of archival tasks and objectives necessary for success.

Approach

The AIMS project had a broad scope but a clear approach. From the outset, the partners realized that the framework would need to acknowledge established practices and infrastructure within archive institutions for managing paper-based collections, the existence of hybrid collections (those consisting of both digital and paper-based materials), and the existence of legacy material transferred in the past and still stored on donors' physical storage media.

As a multi-institutional and multi-functional partnership, the group included archivists, digital archivists, technical developers and repository managers, and other stakeholders within each partner institution. Each of the four institutions have different strengths, different collection specializations, and face different challenges. They vary in size, resources, and capacity, both in terms of parent organizations and archive/manuscript departments within the larger library function. This diversity forced the teams to be flexible, to explore a variety of options, and to compare and evaluate options. The result was a series of collective decisions and a framework that is by its nature not institution-specific.

The project also combined two different organizational models for archive/manuscript departments. The first model is found in larger organizations, where functions of collection development, cataloguing and provision of access are undertaken by different members or groups of staff. This enables (and indeed requires) policies and practices within each function to be well developed and documented. However, there is little continuity of stewardship for a single collection across its lifecycle within the institution. The separate functions may have different priorities, objectives, or ways of working. In the UK, this model is relatively rare outside the larger national institutions, while in the US larger institutions (including the academic institutions among the AIMS partners) are more common and therefore the organizational models of these larger institutions tend to dominate professional discourse.

More prevalent among smaller institutions in both countries are smaller professional staffs who undertake (to a greater or lesser extent) all the functions of collection development, cataloguing, and access, perhaps specializing in a particular subject area. This does give greater continuity of stewardship; however, in some cases, there are fewer resources (and perhaps less pressure) to develop detailed processes and policies for stewardship.

The transatlantic nature of the collaboration allowed the project to work within the established and evolving digital archive communities of both nations, broadened its perspectives as well as its potential audience, and also shaped its methodology. One constant was the presence of legacy collections and anomalies. The collection-focused nature of the project solidified these areas of overlap, resulting in an approachable and accessible framework. All four partners are university libraries or archives, all linked to professional colleagues and networks in other sectors within their national or regional context.

To further ensure its broad applicability, the partners agreed that the stewardship framework should be developed in compliance with established standards, models, and terminology — whether based on archival, technical, legal, or ethical standards. Two standards of note are the Encoded Archival Description (EAD) and the Open Archival Information System (OAIS).

The partners also sought to incorporate existing tools and services, such as Pronom and DROID, and, when possible, to rely on software agnostic or open-source solutions. The University of Virginia, Stanford University, and the University of Hull's collaboration on the Hydra Project[1] prompted a natural choice to use the Fedora-based repository environment. Nonetheless, the Framework does not rely on any particular system; in fact, project partners developed functional requirements for new tools to fulfill the archival functions of arrangement and description.

[1] http://projecthydra.org

So that born-digital stewardship could be completely integrated with systems and processes for its paper-based predecessors, the partners sought to recognize established archiving tools, such as Archivists' Toolkit (AT) in the US and Axiell CALM in the UK. These tools are not explicitly referred to in the Framework, but information detailing the use of these tools by individual AIMS partners can be found throughout the text.

In addition to the development of tools, the project sought to draw upon the significant body of developing initiatives focused on the stewardship of born-digital archives including the following:

Paradigm

http://www.paradigm.ac.ukl

The Personal Archives Accessible in Digital Media or PARADIGM project (2005-2007) was a collaboration between the research libraries of the Universities of Oxford and Manchester to "explore the issues involved in preserving digital private papers through gaining practical experience in accessioning and ingesting digital private papers into digital repositories, and processing these in line with archival and digital preservation requirements." PARADIGM created a workbook documenting their recommended best practices. The PARADIGM project's influence is substantial and further discussion of the parallels and differences between AIMS and PARADIGM are explored in the *Introduction to the AIMS Framework*.

futureArch

http://www.bodleian.ox.ac.uk/beam/projects/futurearch

Also funded by the Andrew W. Mellon Foundation, futureArch at the Bodleian Library seeks "to transform our capacity for working with born-digital & hybrid archives." In particular, Bodleian Electronic Archives and Manuscripts (BEAM) has been working on digital preservation infrastructure, researcher interfaces for hybrid archives and curatorial practices.

Archivematica

http://archivematica.org/wiki

Archivematica is a "comprehensive digital preservation system" offered as an open-source software solution. Based on the OAIS functional model, Archivematica uses a micro-services approach to create an integrated suite of tools for processing digital objects from ingest to access.

Digital Lives Research Project

http://www.bl.uk/digital-lives/

Through the Digital Lives Research Project, the British Library explored personal digital collections in the 21st century. The project inspired a Digital Lives Research Conference and the Digital Lives blog. To date, an initial synthesis of the research has also been published.

Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use

http://www.neh.gov/ODH/Default.aspx?tabid=111&id=37

This National Endowment for the Humanities (NEH) Start-Up grant-funded a project examined the management of the born-digital components of three significant collections of literary material. The project whitepaper is available online and explores issues surrounding preservation and access.

Digital Forensics and Born-Digital Content in Cultural Heritage Collections

http://www.clir.org/pubs/abstract/pub149abst.html

This report, commissioned by the Council on Library and Information Resources (CLIR), was published in December of 2010 and explores how digital forensic techniques typically used by the law enforcement and computer security fields can be applied in the stewardship of born-digital collections within cultural heritage institutions.

Salman Rushdie's Digital Life

http://marbl.library.emory.edu/innovations/salman-rushdie

This hybrid digital collection at Emory University's Manuscript, Archives and Rare Books Library (MARBL) provides a model for arrangement, description, and access to born-digital materials. While the work done on this collection may not be practical for all institutions, the exploration of various issues has been very influential.

Practical E-Records

http://e-records.chrisprom.com/

The blog Practical E-Records was created as a result of Fulbright Scholar Chris Prom's work at the Center for Archive and Information Studies (CAIS) at the University of Dundee. The blog "aims to evaluate software and conceptual models that archivists and records managers might use to identify, preserve, and provide access to electronic records." Posts on specific tools and models were helpful to the digital archivists in developing processing workflows.

Project Narrative

The AIMS project was initiated as an extension of the Hydra project partnership between the University of Virginia, Stanford University, and the University of Hull. The addition of Yale University broadened the project by adding a non-Hydra partner. The project's purpose, objectives, and methodology were refined during discussions with the Andrew W. Mellon Foundation. The project began in October 2009 when funding was confirmed.

The first project milestone was the recruitment and hire of a Digital Archivist at each of the four institutions. All four digital archivists were initially appointed to fixed-term contracts. However, two of the four posts have subsequently become permanent (at Stanford and Virginia) and the other two (at Hull and Yale) were filled via a secondment. All four institutions will retain these experienced staff members assembled for this project.

Once the digital archivists were oriented to the technical, organizational, and archival environment of their institution, the project proceeded via two workflows.

First, the Digital Archivists and their colleagues processed the digital collections identified for the AIMS project, many of which were hybrid collections of digital and paper-based materials. The Digital Archivists shared information on all elements of their work: capture and handling procedures; processing methodologies and tools; ethical and archival issues; and issues of discovery and access. Secondly, the entire project team collaboratively developed the AIMS toolset or framework.

Both efforts were informed and influenced by each other and by the digital archive community in the US and the UK. The collaborative work took place via face-to-face and on-line meetings and environments:

- Face-to-face meetings every six months, involving the Digital Archivists, lead archivists, repository managers, and developers within the AIMS team, and other colleagues from the host institution. These meetings were occasionally timed to coincide with Hydra development meetings (once with the full AIMS team and once with the Digital Archivists), to derive maximum benefit from travel expenditure and to enable archivists and technicians to meet face to face
- Conference calls every two weeks for the full AIMS group (as above)
- Conference calls on alternating weeks for Digital Archivists and the AIMS developer
- Regular in-house meetings at each institution
- During the later stages of the project, brief conference calls every week for Lead Archivists
- Collaborative discussions and drafting of working documents

The project team collaborated with others working in this area and with the digital archivist community through the following means:

- A blog, with postings from the AIMS Digital Archivists and from guest bloggers (for more information, see *Appendix I.1*):
  http://born-digital-archives.blogspot.com/
- Several in-person meetings and collaborative events including:
  - An Unconference in Charlottesville in May 2011
  - A symposium in London in June 2011
  - A half-day workshop prior to and a presentation during the 2011 Society of American Archivists (SAA) Annual Meeting entitled "CREW: Collecting Repositories and E-records Workshop." Detailed accounts of these events are in *Appendix I.2*.
- The creation of the Day of Digital Archives project and blog (see *Appendix I.3*):
  http://dayofdigitalarchives.blogspot.com/

The AIMS Framework was developed progressively through each of these meetings, events, and calls, with objectives being agreed to at each stage before building the next level of granularity. The first task: reaching consensus on the scope, purpose, and definitions of the main archival activities — as referred to in the Framework. For each stage or activity, key objectives (described in archival terms with specific reference to born-digital material) were identified and parsed into decision points and tasks.

With these functions more fully characterized, it was possible to investigate resources and tools. Commencing with a review of existing options (either tools developed specifically for archival use, or those with another primary purpose — for example forensic investigation), tools and software were then tested through real-life implementation with a sample corpus of material from within the AIMS collections. The testing and evaluation focused on the extent to which the tool fulfilled the defined archival requirements such as ensuring authenticity and integrity, and/or documenting an audit trail. While several tools fulfilled some required needs, no single, open-source solution was identified for arrangement and description. In addition, some of the commercial tools tested were not designed for the archival market and required adaptation for archival workflows.

As a result of this unfulfilled quest, the team authored functional requirements for a tool to fill this gap in born-digital archive stewardship. These functional requirements are described more in *Chapter 2: Accessioning* and

more fully in *Appendix H.1*. In line with our general research methodology, this work translates traditional archival principles and practices into a born-digital context.

Lessons Learned

The most basic assumptions were constantly tested during the project. Three formidable challenges were the iterative nature of the project, varying institutional perspectives, and differences in terminology for similar concepts among project partners.

Iterative Processes

Once processing of the project collections commenced, it became apparent that the workflows would have to be iterative both within one archival function and between functions. A closer and more granular definition of archival activities revealed the extent to which they are carried out at different stages in the workflow, depending on individual collections and circumstances. Some tasks must be carried out at a specific place or order in the workflow, while others are relevant to all or can be done at different points. In some cases the deciding factor was archival, sometimes practical or technical, sometimes ethical.

The iterative nature of archival workflows has relatively few implications for the successful preservation of paper-based archives. A suitable physical storage environment is the single most important factor and is relatively easy to define and monitor. With born-digital material there is a greater need to understand, analyze, and assess the implications of decisions made at a particular stage of the workflow to avoid problems or conflicts later. The workflow then must be seen as a whole even when embarking on first steps.

The iterative nature of processing collections at each institution also demonstrated the need for scalability. In particular, accessioning and processing workflows need to allow for and enable digital materials to be transferred to managed storage as soon as possible to ensure preservation of bitstreams. This requires a workflow as free as possible of bottlenecks and labor-intensive processes that prevent this early and successful transfer.

Institutional Perspectives

The second challenge to the AIMS project was the diversity of institutional perspectives. Although this diversity was eventually perceived by the partners as a benefit in building the Framework, it also meant that no single approach, set of assumptions, or workflow steps could be adopted by default. Each had to be defined, shared, and mapped onto those of the other partner institutions so that generic tasks and objectives could be defined for the Framework.

Terminology

The third challenge was language and terminology. The differences both in use and understanding of terminology between the US and the UK as well as between the archival profession and the digital library world of both countries prompted questions and, in many instances, prevented the acceptance of assumed definitions and understandings. Adding to this challenge was the redefining of traditional archival terms to a born-digital context. The partners recognized that, despite differences in terminology, the fundamental archival objectives and outcomes required redefinition of the nature of the activities and tasks required to achieve them. To aid in disambiguating these terms, the project partners created a glossary, included in *Appendix A*.

Conclusion

The AIMS project did not promise to solve all problems associated with born-digital stewardship. In fact, we realized that recommendations could only be for good practice rather than best practice. This is a practical approach but also a recognition that there is no single solution for many of the issues that institutions face when dealing with born-digital collections. Instead, the AIMS project partners developed this framework as a further step towards best practice for the profession.

The AIMS Framework: The Functions of Stewardship

Introduction

One of the primary research outputs of the AIMS project is the AIMS Framework: The Functions of Stewardship. The Framework attempts to map an emerging world combining traditional archival practices with new technologies. While traditional practices evolve (one need only witness the impact of Greene & Meissner's article on "more product, less process" (MPLP) methodology[2] as evidence of this), the increasing sense of urgency at institutions for a scalable methodology of acquiring and processing born-digital materials will change the traditional paradigm even more.

As the four partner institutions worked collaboratively to design procedures for accessioning and processing born-digital materials, we discovered these cannot be isolated activities carried out in one department, or by one staff specialty, or apart from the rest of the archival management workflow. The Functions of Stewardship document the entire lifecycle of born-digital material from the moment the institution becomes interested in acquiring to the instant that a researcher accesses the material.

The Framework is divided into four main functions that should be thought of as sequential steps in a very high-level workflow. However, it is also important to view the process as a whole. Decisions made at the beginning of the process will have a direct impact on later outcomes. Furthermore, with growing legacy collections of data on disks and servers already sitting in our stacks, the process at an individual institution may begin somewhere in the middle or may require moving through the functions in an order different than what is presented here.

The AIMS partners reached consensus that the activities described in the framework are necessary for ensuring the successful management of born-digital and hybrid collections. As described in the *Foreword*, one of the strengths of the project is the diversity of archival environments and practices at each of the AIMS partners' institutions. This diversity, while high level, prompts the AIMS Framework to provide a sound basis for developing more robust and sophisticated local practice.

The four Functions of Stewardship outlined in the rest of this document are:

[2] Greene, M. A., & Meissner, D. (2005). More product, less process: Revamping traditional archival processing. *American Archivist*, 68(2), 208-263.

- **Collection Development:** the actions and policies of an institution to acquire material for end-users as they define them — both current and future. Collection development activities form the basis for subsequent actions and decisions undertaken by the institution as they accept stewardship for and legal ownership of materials from a donor, creator, or seller. This is particularly important as institutions develop their strategies for dealing with born-digital materials.
- **Accessioning:** a core function of archives, wherein an archival institution takes physical and legal custody of a group of records from a donor and documents the transfer in a register or other representation of the institution's holdings.
- **Arrangement and Description:** the processes undertaken by an institution to establish intellectual control of the material following the physical control secured during accessioning. It also prepares the material for discovery by preserving the context of the materials, and prepares for access by applying appropriate restrictions.
- **Discovery and Access:** the systems and workflows that make material, and the metadata that support it, available to users while ensuring compliance with any access restrictions with. The process of discovery and access requires some action on the part of individual users — for example carrying out a search or requesting an item.

Each functional area is further described in this document with necessary objectives identified for each. These objectives are further detailed through expected outcomes, decision points, and tasks. In addition, "keys to success," or areas that should be addressed and conditions that should be put into place before beginning work in an area, are defined for each objective.

One area intentionally not addressed in this project is digital preservation — the specific practices developed to ensure the long-term viability and security of data. The reason was twofold. First, while an emerging discipline, digital preservation has many well-documented best practice models and methodologies. Reiterating what others have already determined would not be useful. Instead, the framework assumes that efforts outside of the archival functions ensure the viability of data. This leads to the second reason digital preservation was not discussed: it is larger than the scope of this project. Digital preservation is a major infrastructure issue for libraries, archives, and other institutions. The only way to achieve reasonable success in digital preservation is in economies of scale wherein the nuts and bolts of preservation (storage space, repository infrastructure, refreshing of media, etc.) are carried out in the same way for all digital content. In this way, the specific archival activities that are explored here do not overlap with preservation activities.

Appraisal is also not defined as a specific, separate function. Rather, appraisal activities are included in any or all of the first three functions within this framework — collection development, accessioning, and arrangement and description. Principles, strategies, and tasks related to appraisal process will appear within each function.

As a final note, the AIMS Framework bears some resemblance to the PARADIGM Workbook in scope and content. However, PARADIGM contains much more detail about acquiring collections and collection development. While more detail is useful, the PARADIGM Workbook sometimes lacks the broad and holistic viewpoint that the AIMS Framework can and will provide. The project partners hope that both PARADIGM and the AIMS Framework can be used together by institutions working towards the establishment of practices for the stewardship of born-digital materials.

1. Collection Development

Definition and Scope

Collection Development: the actions and policies of an institution to acquire material for end-users as they define them — both current and future. Collection development activities form the basis for subsequent actions and decisions undertaken by the institution as they accept stewardship for and legal ownership of materials from a donor, creator, or seller. This is particularly important as institutions develop their strategies for dealing with born-digital materials.

Preface

In the initial stages of the AIMS project, the activities described in this section of the AIMS Framework were referred to as "pre-accessioning intervention." From an archivist's perspective, this is a more accurate description than "collection development" because it highlights attempts to determine what actions should be undertaken or what information should be gathered in order to lay a solid foundation for the long-term stewardship of born-digital archives in order to inform and assist curators when working with donors during this early stage. However, since this work is undertaken within a larger framework across disciplines and with a variety of other archive or library staff, the term "collection development" is more universally understood.

Until recently, born-digital materials were often viewed as an adjunct to the paper or analog materials in a collection. They were seen as less important, in many cases thought to be duplicative or uninteresting, and perhaps as items that could be discarded. Specific collecting activities related to born-digital materials within manuscript collections were sporadic and undefined. Ensuring preservation and access usually included printing out the digital files. While feasible when there are only a few items, this activity is neither sustainable in the long term nor preferable. Despite their complications, born-digital materials are more flexible, enabling full-text search or other interactivity. The loss of this flexibility downstream in a discovery environment has led to a growing effort to keep digital files (rather than printing or discarding them), whether or not they are duplicates of analog material.

Traditionally the selection and cultivation of a collection has been the sole purview of a subject curator or archivist, possibly working in conjunction with an acquisitions committee. While the processing team or archivists may have been called on to assist with parts of the process, communication during this initial phase might be limited. When dealing with born-digital materials, however, it is best to employ a more collaborative approach from the outset: technical expertise and experience with newly designed workflows (or those just being tested) from archival or digital staff will aid the curator in appraising materials, performing test captures, and identifying any issues related to accessioning, processing, preservation or delivery. There are numerous examples of scenarios where technical and legal expertise would be necessary, including: undertaking research

on new capture methodologies from a media type not previously encountered; negotiating permission to capture or extract data from a proprietary web service; assessing the feasibility of taking material dependent on software or other programs that require significant commitment to deliver or render; and understanding the licensing and intellectual property rights implications of capturing or copying software as well as data. These activities are substantially different from those undertaken when dealing with analog materials, and it is best to discuss these issues with a team from the outset.

The team approach in the collection development phase will allow all parties to:

1. be aware of broad issues as they arise in order to develop strategies for incorporating them into current and future workflows,
2. work closely together to better understand the institution's ability or capacity to receive the digital materials in question and to undertake long-term stewardship
3. have a full understanding of the implications of donation, acquisition, processing and delivery.

Keys to Success

Collection development is the first step in the AIMS Framework for born-digital stewardship. Decisions made regarding born-digital materials by the curator and donor will affect each additional step of the archival process and the long-term plan for accessibility. While the objectives below differ slightly from the traditional collection development experience, they serve many of the same functions, including establishing trust with donors and depositors and creating comprehensive documentation for future activities.

Before embarking on these objectives, institutions should also foster discussions about the policies and procedures highlighted below. These discussions will require consideration of future scenarios, an activity made all the more difficult by rapidly changing technology and professional practice. The methodology or practice at each institution will differ in the approach to developing policies and procedures. An understanding of your institution's strategic environment and risk tolerance will be essential in successfully navigating these decisions. Some institutions will require deliberation and the creation of policies as a first step; others are more likely to favor experimental action. The goal is for all departments and staff involved to agree on expectations, abilities and capacity. There is no right or wrong way to do it, but waiting for a perfect workflow or tool to evolve may mean losing those materials through data loss or to competing organizations. Spending time at this stage to think through future activities will result in less confusion and difficulty later.

Born-Digital Collecting Policies

The collection development policies of an institution are designed to guide the acquisition of materials according to their mission and collecting goals in order to meet the needs of their end users. The implementation of the collection development policy in relation to born-digital materials might involve:

- prioritizing collections based on the needs or strategies of the institution and its user communities (stated or implicit)
- developing relationships with donors

- assessing born-digital and analog materials and the relationship between them
- evaluating how the former might fit into current technological strategies or push further development at the institution

Therefore an institution's born-digital collecting policy needs to establish the institution's position — its principles and general standpoint — on a wide range of issues which have implications for stewardship. This will ensure that it is effective in guiding discussions and decisions relating to specific donations and individual accessions during collection development activities. These discussions will determine how an institution's policy is applied in a specific instance, any exceptions to the policy, as well as how options within it will be recorded in a legal agreement.

A born-digital collection development policy should supplement an institution's general collecting policy, and include information about:

- Method(s) used for transfer and/or capture of materials
- Methods for identifying and dealing with files that contain viruses or other threats to preservation
- Options for dealing with files that are duplicates, redundant, or out-of-scope
- Criteria for capture (or acquisition by other means) of proprietary or open-source software, or of hardware
- Strategies and methods for preserving materials (what is preserved and how)
- Strategies and methods for providing access to materials (what is delivered to the user and how)
- Policies and strategies for dealing with confidential or other sensitive content
- Conditions governing access (restrictions, limits on access, users) and how they are applied and enforced
- Policies relating to intellectual property rights, including Creative Commons Licensing and copyright (the role of the institution and how it is undertaken)
- Retention (or not) of original storage media
- Categories of digital material (AV, databases, text, etc.) which the institution is able to preserve, manage, and deliver, with indicative listing of file types and formats, and limitations where applicable
- Methods for ensuring and demonstrating integrity and authenticity, with associated criteria. (As discussed in *Chapter 3: Arrangement and Description*, more development is needed in this area.)

These issues have technical, archival, ethical, and legal elements. Many of them relate to the technical processes required for accessioning and delivery of materials and are discussed more fully elsewhere in this document. These processes have ethical and legal implications that the donor needs to be aware of and to understand in order to give informed consent. For example, if the institution's default policy is for a 'bit-for-bit' capture,[3] will the donor be asked for their preference? How would the institution handle files previously deleted by the donor/creator but which are included in the data capture? Will there be a difference between what is captured and what is accessible to the public (for example, when is the bit-for-bit copy retained for preservation and

[3] As an example, see *Appendix F.7*: Beinecke Rare Book and Manuscript Library's Born Digital Archival Acquisition Collection & Accession Guidelines, specifically – "In acquiring born digital materials … the capture by "snapshot" of all working files on a specific computer, will be the preferred method of acquisition; in most cases BRBL will wish to capture entire digital environments without any advanced collection editing by creator or curator."

processing purposes only and not for access)? Will material available online differ from what's available on-site (for example via a standalone computer in the reading room)? Will legacy/transfer storage media be retained and what software or hardware will be captured or acquired? Software may be needed in order to render the files with their significant properties[4] but capture of proprietary software from a donor may contravene licensing agreements.

There are several useful papers discussing the issues and ethics of working with born-digital materials.[5] However, your institution should have a written statement that the curator may use as a primary point of reference.

Technical limitations at an institution might initiate a list of preferred formats based on capacity and ability.[6] This may be driven by preservation strategies and, to a lesser degree, the current capability for delivery. However, the acquisition of born-digital material should be based firstly on a curatorial appraisal of its fit within the collection development policies of the institution. In addition, a feasibility study or technical appraisal should be performed by archivists and/or technical specialists before final decisions are made. While an institution might take in a format that is not on its preferred list, it would need to understand that it cannot guarantee the same level of stewardship —i.e., preservation might be only at the bit-level.

Recognizing your institution's ability and willingness to collect born-digital material and defining the parameters of this effort is key. Many institutions are redefining their collecting policies and overall strategies for the 21st century, and born-digital is recognizably a huge issue for many repositories as was demonstrated in the 2011 OCLC survey report.[7]

Digital preservation strategy

An institution must have a good understanding of its technological capabilities and must have some sort of preservation strategy in place or in development in order to undertake stewardship of born-digital archives responsibly. This strategy must include the management of material from the moment of transfer, through processing in a virtual workspace, and finally ingestion into preservation and/or delivery repositories. A complete understanding and description of the infrastructure will include:

- storage environment
- equipment for transfer, capture, and quarantine

---

[4] For example, at Stanford, when the Peter Koch computer files were acquired by a logical capture, the fonts associated with his In-Design and Quark design files were not captured. This created an inability to render the printer's designs accurately in the virtual machine – especially as many of the fonts were no longer available.

[5] For example: Matthew Kirschenbaum, Richard Ovenden, and Gabriela Redwine, *Digital Forensics and Born-Digital Content in Cultural Heritage Collections* (Washington: Council on Library and Information Resources, December 2010); and Digital Lives Research Project (http://www.bl.uk/digital-lives/)

[6] For example: Deep Blue Preservation and Format Support Policy at the University of Michigan (http://deepblue.lib.umich.edu/about/deepbluepreservation.jsp) and Wellcome Library Digital Curation toolbox (http://library.wellcome.ac.uk/node289.html)

[7] The British Library's website discusses their collecting policies for the 21st century (http://pressandpolicy.bl.uk/content/default.aspx?NewsAreaId=312) and the anxiety on the part of archivists for dealing with born-digital materials is documented in "Taking Our Pulse: The OCLC Research Survey of Special Collections and Archives," Jackie M. Dooley and Katherine Luce, OCLC Research, Oct. 2010.

- maintenance activities; personnel and skills required
- planning and communication strategies

Not all institutions need build their own digital repository. An acceptable strategy might include joining a local or regional repository.[8] Wherever the location of the specific repository, institutions need to ensure and demonstrate that they can and will undertake responsible stewardship, or question whether they should in fact be collecting born-digital materials.

Legal agreements

Many institutions will already have in place a template for agreements with donors (or depositors, sellers, or vendors) which covers analog materials. Before active collection begins, the agreements should be amended to cover born-digital materials. As with the collection development and preservation policies above, the agreement should acknowledge and make explicit reference to the salient characteristics of born-digital materials and the additional issues which arise with born-digital or hybrid archives. This will facilitate common understanding between donor and institution, ensure informed consent, and record key decisions and information for future reference.

Donors and Trust: University of Hull

Simon Wilson
Digital Archivist, University of Hull

An organization recently contacted the Hull History Centre regarding the transfer of over 100 linear metres of their historical archives, dating back over 170 years. The organisation also expressed a willingness in principle to participate in the AIMS project.

Despite several meetings to discuss the potential type and range of born-digital material to transfer, the organization was hesitant and eventually withdrew from the project. Having recently transferred their paper archives to the Hull History Centre, there was an on-going relationship with the donor and a level of trust and understanding about the value and importance of archives. The organization recognised that the paper archives had historical value, even though they were clearly no longer using the paper-based records on a daily basis and could not justify the space those materials occupied in their office. Although Hull History Centre staff were thinking about the continuity of records from paper to born-digital, the organization regarded the records differently and had not thought about how they would continue preserving the legacy of their work in the digital age. The organization's digital material was still actively being created and used, and less likely to be perceived as being "archival" or having historical value in the same way that the paper archives clearly did. The born-digital files accumulating on the servers were less visible than their paper predecessors, and shortage of space was less of a concern. Data security and the possibility that sensitive material would be transferred was also a worry — although voiced somewhat vaguely, perhaps because the risk would only become relevant and specific once the principle of transferring the born-digital material was accepted.

The lesson learned from this experience: place greater emphasis in the initial discussions with potential donors on the continuation of established practices relating to material of archival value, whatever its format, rather than on the format of the material. The reluctance of this organization to discuss with Hull History Centre staff the born-digital material that they were creating hampered the Centre's ability to identify or recommend possible material for the archives, or to offer reassurance about the protection of sensitive material. This uncertainty also led to the organization's concern that the identification of material for transfer would take considerable time and effort on their part.

Also evident was the fact that, in the future, the Centre must be clearer in explaining that the nature of born-digital archives necessitates the capture of electronic records soon after creation — much sooner than is traditionally the case for a paper-based materials. In developing a user interface for born-digital archives, the Centre will actively look to demonstrate to donors the ability to safely store and control user access for

[8] See the California Digital Library (www.cdlib.org) or the Orange Grove Repository, Florida's regional digital repository (http://www.theorangegrove.org/OGMain.asp)

At a minimum, the agreement should include the following elements:

- Scope and description of materials being transferred, both analog and digital, in either aggregate or particulars. This may make reference to a survey compiled of the material, of the donor's working or digital environment and other related information
- Processes for reporting and documenting acknowledgement of successful receipt/capture
- Arrangements for transfer or capture of born-digital materials — both time frame and methodology
- Implications of capture method and associated requirements (for example, how files previously deleted by the creator or donor — but recovered during capture — are to be dealt with)
- Reference to preservation of digital materials (what is preserved, what is promised or guaranteed, any caveats or limitations). For example, the institution may explicitly exclude any obligation to meet the requirements for admissibility of born-digital material as evidence in court.
- Conditions of or limitations on access (for example, online or on-site, open to all to institution's community, to specific users[9] or IP addresses, etc.); types of material to be restricted because of confidentiality, data protection (in the UK) or other legal or ethical factors; how these are to be identified or defined; types of delivery
- Ownership of materials (relating to analog material, where content is unique) or exclusivity (relating to born-digital materials, where content is a copy)

Donors and Trust: University of Hull — continued from previous page.

materials stored in the Centre's digital repository and allay any fears that may arise from phenomenon such as wiki-leaks.

Interestingly, another organization — which has been regularly transferring material to us since the 1960s and which was equally reluctant to include born digital records in these transfers — has recently itself raised the issue of its born-digital archives. Early reluctance was the result of concerns about access to and misuse of the material, from the viewpoint of intellectual property and reputation rather than personal confidentiality. A change of staff within the organisation, together with a greater emphasis on preserving a record of its more recent activities, has helped to overcome those initial reservations. The risks of transfer and more open access are still present, but the organization is now able to see the advantages of creating a born-digital archives, as well as the obstacles to be overcome.

Workflow

Finally, before beginning collection development actions, an institution should ensure that the archival and curatorial staff has an understanding of the workflow for born-digital materials and the delivery possibilities currently available. Even if the workflow is still under development, all parties should be aware of what the current plans are for management of the material. As discussions between the curator and the creator or donor progress, archival and technical staff should be consulted so that they may contribute to analysis and decisions made and should also be kept up to date regarding incoming collections so that they may plan for storage requirements and other collection needs.

Objectives

OBJECTIVE 1. Establish relationship with donor

Outcome: *A collaborative relationship is established between the curator or institution and the donor, the*

[9] In the UK allowing access to content to some users (or classes of users) but not others may be required in the case of records covered by the Data Protection Act. However, in other cases, under the Freedom of Information Act, once closed files are opened for an individual (e.g., a preferred researcher working on somebody's personal papers) that material is considered to be open to everybody.

contents, formats, and requirements of the collection are identified, enabling the curator to determine that the material aligns with institutional collection development policies.

In the realm of traditional collecting, establishing a relationship with a donor is often handled by a subject curator. This is an area that is fairly well defined in the Paradigm Project's website.[10] The conversation with a donor of born-digital materials must at least establish an agreement on what is to be donated (or a part of it, in the case of hybrid collections).

From the donor's perspective, there are topics that should be considered in order to establish trust in the institution's ability to handle their material and meet other requirements. What are the long term preservation and migration capabilities and plans? Does the donor require a copy of the data captured as a back-up or for reference? Can the institution provide access to the material in a meaningful way? Will restricted files be protected and inaccessible until the specified date — and how will the institution assure this? Provenance or chain-of-custody issues will also be important to ensure and to demonstrate that the files have not been tampered with prior to transfer and are not affected by the transfer process itself.[11] All of these issues should be discussed and documented. The archivists and technical staff need to lend expertise and support to the curatorial staff so that they are aware of the institutions' capacity and workflows. This will ensure that the institution is able to take in and steward the materials they acquire.

Documenting this information will not only aid the donor, who may not be familiar with or comfortable navigating or discussing digital concepts and processes. It will also help the curator define the scope of the collection and determine when and if to involve other staff in the process. Discussions between the creator or donor and the curator, with assistance from a digital archivist, will include timing and methods for transfer and capture of data, possibilities of data corruption or loss, processes for acknowledging what was captured, plans for long-term preservation, and possible processing or delivery strategies. The AIMS project team developed a digital survey (see *Appendix F.1*), based loosely the work of the Paradigm Project, to serve as a prompt in meetings to elicit information about to donor's born-digital materials. The questions are designed to frame a dialogue to purposefully uncover personal and digital work habits, the extent of material, formats and locations, passwords, use of peripherals, etc. Information on migration of material, software changes, back-up strategies, and the relationship between digital and analog material is also very useful.

While it is essential for the digital curation team to document how and where a creator worked in the digital arena, a curator may desire to record information about their work and workspace in other ways. The British Library has developed an initiative called "enhanced curation" (see *"Enhanced Curation at the British Library" on pg. 12*) which might consist of video interviews or high-resolution photographs of a donor's workspace.

All of the data collected during this period will inform the creation of the legal agreement later between the donor and the institution. It will also set up reasonable expectations for both parties on what to expect once the data is transferred. One other consideration would be to discuss issues of transfer after their death. Digital

[10] http://www.paradigm.ac.uk/workbook/record-creators/nature-relationship.html

[11] The Beinecke's collection and accession guidelines in *Appendix F.7* address this last issue.

wills are becoming more common and this, along with plans for depositing additional analog material, may be a topic that a curator feels comfortable discussing with a donor.

Decision Points

There are many decisions within this objective that are to be made by the donor as well as the institution, within the context of collaborative discussions. These discussions will assist institutional staff (curators and archivists) in determining the desirability of acquiring the born-digital materials offered. In addition to establishing a relationship with the donor, the curator needs to document and justify that the subject focus of the collection aligns with the institution's policies, priorities, and interests. The curator also must to decide if the content is sufficiently complete and significant to merit the initial work and cost of transfer as well as the on-going commitment to preserve and make the materials accessible. How does the interrelationship of analog content (if any) and born-digital materials inform decision-making about the born-digital content under consideration? For example, the curator might decide that duplication of analog and digital is permissible (as it is unlikely to be sustainable to compare materials in the two formats) or that the digital files are only a backup for the "printed" copies.

Staff may also take an initial view as to whether the character and scope of the born-digital material lies within the bounds of the institution's ability to receive and properly steward it. However, a full determination of viability is the objective of the analysis and feasibility study, which is the next stage of the process. If the donor's terms (as understood at this stage) appear to be acceptable to the institution at this point, a legal agreement could be drafted. Yet it is important to recognize that an alternative outcome is possible if something uncovered during the survey or background research precludes moving forward to the next steps.

Tasks

The information gathered during discussions with the donor — from the survey or other research — should be reviewed according to the various criteria within collecting policies and guidelines (implicit or explicit). A report could be produced (and may be required by an acquisition committee) justifying that a collection and its contents fall within relevant parameters. At this stage it is also useful to scope (or at least flag up) any hidden costs or other issues that may arise, or technical assistance that might be required. Factors relevant here are:

- the formats and extent of born-digital material in the collection
- the relationship or overlap between born-digital and analog material
- the creator's work habits, use, platforms, and software
- the likelihood of future transfers or captures and their frequency
- options for data transfer or capture for current materials and future accruals
- in the case of capture of live or active files, how the scope of current content and future accruals will be defined
- any requirements on the part of the creator or donor for ongoing access to content and on what terms
- any requirements on the part of the creator or donor for restrictions on access to some or all content and the timeframe for these
- any legal or other restrictions or conditions, outside the donor's or institution's control

If enhanced curation methods would be beneficial to the institution or collection, these should be discussed at this stage. While this work is often curatorially driven, archivists might be involved in carrying out or assisting with some of the activities and in accessioning the resulting material as part of the collection. Tasks might include:

- documenting physical media, workspace or storage space through photography
- recording an interview (audio or video) with the creator or donor about their use of technology and their 'digital life.'

OBJECTIVE 2. Analysis and Feasibility Study

> ***Outcome:*** *A determination is made as to whether the collection can be reasonably acquired, managed, and preserved within the constraints of the institution's resources.*

Once the institution is certain it is interested in acquiring a collection, they must determine if they are technically capable of acquiring and managing it. This process includes an assessment of the nature of the material, the costs of the activities to steward them, and the resources available to the institution (including staff, budget, and time). The collection development policy again becomes important in determining the degree of cost the institution is willing to incur to acquire the material. If the collection is highly valued based on the priorities set out in the policy, the institution may be very willing to put a large amount of resources into developing the technological infrastructure to handle the material. If it is weakly aligned with collecting policies, they may not want to invest much at all.

Costs associated with the stewardship of born-digital resources are not well understood or documented.[12] While some costs associated with traditional archival practice will be similar, it is more complex in the digital world. Costs need to be considered for building or purchasing equipment or for external services to properly accession materials and to process them. Substantial costs may also be incurred in training and/or hiring staff to undertake technical work, to research new tools, to build a born-digital curation team in-house, and to develop collaborations outside of the institution. The appraisal itself may require the use or development of new tools and methods, meaning that the feasibility analysis itself incurs costs.

Decision Points

It may not be possible to identify all the requirements or issues that will arise during the different stages and activities of born-digital stewardship. However answering some key questions will help to inform the decision making process. Do file types or formats within the collection correspond with those which the institution has already accessioned and managed, or has determined that it is ready and able to do so? What is the likelihood or possibility that content or metadata is corrupted, unstable, unreliable, or incomplete? Does any content require interoperability with data or tools not present or that cannot be made available to the institution? Does any content require specialist software or a specific platform environment to be rendered fully intelligible for documentation and research purposes? If so, is the institution prepared to commit to paying software license

[12] The LIFE project, a collaboration between University College London (UCL) and the British Library is investigating costs associated with preservation (although this will not cover all costs of stewardship). http://www.life.ac.uk/blog/category/digital-preservation/

costs or preserving the original platform functionality? Over what period of time? Is the institution prepared to commit to making data available to users within that platform environment? What is the institution's general policy regarding preservation of original storage media? Is there a reason in this particular case to depart from that general policy?

Tasks

- Assess institutional resources available, including technology, staffing, and funding.
- Assess file types and formats of digital content against preferences and exclusions for file type and format established by the institution; tools such as DROID can be used to produce information on file formats.
- Assess volume of born-digital material relative to storage and management capacities of the institution and its digital repository (if applicable).
- Request an increase in capacity if required and investigate costs.
- Assess the condition or "health" of digital content. The diligence of the creator or donor in activities such as keeping anti-virus software up to date may give clues in this analysis.
- Assess the dependency of content on software or platforms and the cost or other implications of this dependency.
- Determine the views and commitment of both parties on the importance of preserving original storage media as well as contents.
- Determine the practicality and feasibility of ongoing transfers of data over time if these are anticipated.
- Determine the nature of migration or transformation processes (paths, tools and protocols) which will be required (for example disk imaging) and their implications.

Appraising and analyzing content is essential during the collection development process although tools to do this effectively are few at this time. Most of

Enhanced Curation at the British Library

Rachel Foss & Jeremy Leighton John
Personal Digital Manuscripts Project,
British Library

The British Library's enhanced curation initiative grew out of the AHRC-funded Digital Lives Research Project, led by the British Library along with University College London and the University of Bristol. Running from 2007 until 2009, this project focused on personal digital archives and their relationship with research repositories. It became clear in conversations with users that the research value of digital objects could be significantly increased by the collection of their contextual information. This recognition led to the Library's enhanced curation work: taking the opportunity to engage further with living creators at the point of acquisition to create extra content recording as many aspects of their work and environment as possible, and providing an additional resource for researchers to use alongside the material which constitute the archives per se.

Within the Library's literary and scientific collections, this extra content has so far included using digital photography to record a virtual panorama of writers' and scientists' workplaces (studio and laboratory), recording interviews with creators where they give a retrospective context to the material we are collecting for the archives, video conversational tours of their habitats, and photographic capture of material (such as a personal library) that is not normally within the scope of the British Library's manuscripts collection policy.

Enhanced curation has also been used as a way to record the acquisitions process itself. For example, in 2009, we created an audiobook diary record of the acquisition of the archives of John Berger, which was later used both as part of an initial promotion of the acquisition and as an enhancement of the Library's catalogue record (http://audioboo.fm/britishlibrary?page=1). Collections which have so far been included in this initiative include the archives of Ted Hughes, Wendy Cope, Anne McClaren, Donald Michie and James Lovelock.

the collections tested in the AIMS project were legacy collections and the partner institutions will continue to do more testing and development in this area. Tools such as Forensic Toolkit, DROID and Karen's Directory Printer, used by AIMS partners during accessioning processes, may also be relevant to appraisal and analysis during collection development (See *Appendix G* for technical reviews).

OBJECTIVE 3. Negotiation & Agreement

> ***Outcome:*** *The informed consent of both institution and donor is formally documented in legally binding agreements including a gift or purchase agreement.*

An institution's subsequent actions of stewardship are based on decisions documented in the signed legal agreements that originated from the initial discussions between the two main parties: the donor and the curator. Therefore these documents are one of the most important products of the collection development process. They provide the roadmap for future work, ensure and demonstrate comprehensive and clear understanding of both parties, and are a legally binding document for both parties.

Decision points

Not all questions or issues raised will have answers at this stage and so it may not be possible to record a decision or agreement. A seemingly simple discussion about sensitive or restricted material in the analog world raises issues on multiple fronts in the born-digital world. It is not as easy to look at the "documents" — i.e., the files — themselves without a viewer. Issues such as the identification and segregation of files to be closed for a period of time or to be returned or deleted may present the archivists and developers with a new problem set. Each new format will invite discussions or require researching or testing of new tools and workflows. It is therefore important make and to document decisions relating not only to precise actions or methods, but also relating to the general principles or strategies that will be applied if specifics are not yet known.

Tasks

- All decisions and information discussed here should be written into the agreement, which may be supplemented by correspondence between the curator and the creator or donor and/or by institutional policies referred to in the agreement or correspondence. Wherever possible all decisions should also be recorded in the institution's collections management system, whether tracked in paper or electronically (see also *Appendix F.5*: Guidelines for Creating Agreements at Stanford University).
- Among the issues documented should be: the extent of institution's undertaking to receive, preserve, process and deliver the materials; managing restrictions and access; the rights and requirements of both the creator or donor and the institution; and the commitment of both parties to uphold rights and meet requirements. A comprehensive template for legal agreements, as described above, is a crucial tool in creating this documentation and in ensuring that it, in turn, is comprehensive.

OBJECTIVE 4. Prepare for Accessioning

> ***Outcome:*** *The agreements are finalized and documented and the transfer and immediate storage of all material has been planned.*

Once the legal agreement is finalized and signed, the final steps to prepare for accessioning may be undertaken.

This objective includes the activities needed to set up the physical transfer of custody for the collection and the "unpacking" of content from transfer storage media. This might involve accessioning of new material in active collections or refer to retrospective accessioning of content from legacy collections which remains on its original storage media. The next section will discuss this distinction in greater detail.

Decision Points

A checklist will help to ensure that the institution has completed the necessary preparatory steps. These will include:

- documenting that the legal agreements are signed by all parties and in hand
- that any outstanding technical issues have been discussed and documented
- that appropriate staff and data storage capacity required to handle the transfer are in place

Tasks

- Arrangements, methods, and timelines for transfer and capture should be finalized, agreed, and coordinated with analog material (if appropriate). This should take into account resources needed, whether of staff, equipment, or time.
- In some cases, it may be useful to perform a test capture to check assumptions and allow problems to be identified in advance. This may also be an appropriate time to carry out any enhanced curation techniques such as photographing the workspace of the creator or recording an interview.
- Finally, documentation should be updated and shared with archival and technical staff as appropriate.

2. Accessioning

Definition and Scope

> ***Accessioning:*** a core function of archives, wherein an archival institution takes physical and legal custody of a group of records from a donor and documents the transfer in a register or other representation of the institution's holdings.[13] Accessioning has four main functions: physical and administrative transfer of records; review of general content and condition of records; creation of initial control tools; and assessment of future needs for arrangement, description, and preservation.[14] These functions serve as the basis for the objectives of the accessioning function within the AIMS model.

Preface

Good archival practice entails accessioning material as soon as possible when acquired by the institution and for resourcing this task as a priority. Accessioning takes legal and administrative custody of the materials with minimal risk to clarity of provenance or authenticity being diminished over time. Furthermore, removing threats to preservation is as important to born-digital material as much as paper-based records. In fact, with born-digital the threat of potential deterioration and data loss can occur much more quickly.

Accessioning is the step where the institution begins its management of the collection. The process includes physical and administrative transfer of records; review of the content and condition of the records; creation of initial control such as accession records and documentation; and, finally, assessment of future needs.[15] The process of establishing custody and control over an accession allows the archivist to undertake further appraisal, arrangement, and description of records, which then enables the records to be made available for use. In-depth assessment and documentation processes during accessioning will provide substantial information to colleagues within the repository, as well as to potential researchers. Most critically, accessioning prompts archivists to document necessary restrictions on access, use, or reproduction.

These activities can form a kind of "baseline processing," if necessary, as they give the basic intellectual and administrative control that is most important to the institution's continuing curation of the material. Indeed, some collections, both digital and analog, may not need further work than what is accomplished during this stage. Therefore, a general accessioning policy must specifically address born-digital materials since it may be possible to integrate materials into the collection in a meaningful way at this point, assuming that the resources are in place to deal with various kinds of media and file formats.

---

[13] Accession," in *A Glossary of Archival and Records Terminology* (ed. Richard Pearce-Moses). Chicago: The Society of American Archivists, 2005), http://www2.archivists.org/glossary (last accessed July 12, 2011).

[14] Kathleen M. Roe, *Arranging and Describing Archives and Manuscripts* (Chicago: The Society of American Archivists, 2005), p. 45-56.

[15] Roe, p. 45-56.

Alternatively, the paper or analog portion of a hybrid collection can be accessioned first with the expectation that the born-digital will or might be incorporated later; this has been a standard practice in many institutions to date. A third option may also be to review some media (e.g., if the infrastructure exists and the volume is not too high) so that it can be done at the point of accessioning. No matter which scenario is chosen, the fact is that institutions must be able to accession born-digital materials to the same effect as they do with paper materials in order to take advantage of MPLP[16] practices, and to ensure timely transfer into appropriate storage to ensure preservation.

To be able to carry out accessioning as a baseline level of processing, institutions must develop tools and workflows that enable them to carry out the objectives described throughout this section of the Framework. Accessioning is not a trivial undertaking, and a significant investment is likely to be made in technology and staff training. Success is not impossible however, and guidance on specific tools is given in the technical reviews found in *Appendix G*.

As it is defined in this section, accessioning is an activity carried out with materials in collections that have not been physically processed prior to this occasion. It is very true that many institutions may find themselves in the situation of having a significant amount of legacy materials that have been physically stored on carrier media but not further managed. Indeed, the AIMS partners found themselves in this situation. Many of the objectives described in the remainder of this section are applicable to these legacy materials and will assist institutions in managing them. There are many other issues specific to legacy materials that are not explicitly addressed in this Framework such as the renegotiation of donor agreements, specific issues surrounding fragile or obsolete hardware, as well as workflows for modifying or updating accessioning and processing documentation. This seeming oversight is in fact intended to shift the focus of this document to the practices that will assist institutions in moving forward with new collections. Workflows for dealing with legacy materials should be addressed elsewhere to avoid unnecessarily complicating this emerging best practice model.

Keys to Success

To successfully carry out accessioning processes, several conditions must be met. The first is ensuring that donors have confidence that the desired outcome of secure transfer of appropriate data will be achieved. The donor must understand the policies and processes associated with accessioning sufficiently so that they can participate effectively in the process, providing necessary information and guidance. For example, using the forensic disk image technique during accessioning, which obtains an exact, bit-by-bit copy of the data on a disk or hard drive, can unintentionally allow the archivist to view and recover records or data that the donor does not intend to transfer, such as deleted files that have not been wiped from the system.

Similarly, the archivist and the collecting institution, too, must have confidence that they will be able to gather sufficient information to establish an appropriate level of physical, administrative, and intellectual control over the materials being transferred. A collecting institution should obtain a sufficient level of control over transferred records to allow them to manage and maintain them, both through further processes of

[16] Greene, M. A., & Meissner, D. (2005). More product, less process: Revamping traditional archival processing. *American Archivist*, 68(2), 208-263.

arrangement and description and in terms of ensuring the records' long-term viability. This is essential if the collecting institution is to maintain the integrity and authenticity of the records. These kinds of detailed inventories can be automatically created using tools like Karen's Directory Printer or FTK Imager, even by institutions with smaller staffs or expertise (see technical reviews of these tools in *Appendix G*).

To this end, archivists should establish guidelines appropriate to the types of records, curatorial areas, or record creators for which they and their institution are responsible. Furthermore, there are tradeoffs to be made in the scalability of these processes. If a collecting institution receives large transfers of digital records on media, the archivist will have to make difficult decisions about which parts of the process are necessary to ensure that the records are accessioned in a timely manner. Scalable accessioning workflows are also essential to redefine prioritization of work between paper and unaccessioned or under-accessioned born-digital records.

A second factor contributing to success in accessioning is having both technical knowledge and an infrastructure capable of handling the transfer of electronic records of various kinds, which require various transfer processes. As collecting institutions begin to receive greater amounts of digital records, this will likely include both "obsolete" and recent media formats. While archivists need not have the capacity to deal with all media formats or transfer scenarios within their institution, they must nonetheless be prepared to make a determination of what they can, and cannot, handle. Archivists should therefore establish clear guidelines on what types of media they can easily handle locally and those that might require work with a vendor or another institution. With this in mind, a collecting repository should recognize the budget implications of each case, as in some cases it may be more cost-effective to hire a vendor to handle the reformatting and transfer according to the institution's guidelines — assuming they can provide the information needed to ensure that the integrity and authenticity of materials can be verified.

A third key to success is careful selection as part of an overall collection development policy. As was addressed in *Chapter 1: Collection Development*, without an appropriate policy that addresses selection criteria and legal agreements, the archivist assigned to manage accessioning and other custodial responsibilities may not be able to establish an adequate level of legal custody for the records. For a collecting institution to gain legal custody of a body of records, they must complete a formal legal agreement — a bill of sale or deed of gift — for those records. The terms of transfer and obligations and permissions described in the agreement should originate in negotiations completed with the donor during the collection development phase.

Accessioning also benefits from being carried out as soon as possible after selection, to better ensure preservation and integrity of digital content; if issues are encountered (for example, a virus amongst the files), there is an opportunity to repeat the transfer process. In addition, the collecting repository's organizational knowledge about the donor, the creator, the transfer process, and the records is at its strongest at this time, and therefore is most effective in facilitating and informing subsequent accessioning activity.

To successfully accession digital content the repository must also have in place a "capture policy," or set of guidelines. Such guidelines should specify institutional preferences for the method of data capture (for example forensic, or bit-by-bit, imaging versus logical, or selective, copying), the treatment of physical media after capture is complete, and the handling of unwanted or duplicative files. Other considerations include capture/transfer methods with guidelines for responding to unsuccessful captures, procedures for handling media, and

procedures for working with the short-, medium-, and long-term storage environments to ensure safe submission and documentation of metadata associated with the born-digital content.

Institutional practices and workflows regarding accessioning born-digital materials must also be in place to further guide practice. These include workflow models such as when or if to use accessioning as base-line processing; when (or whether) to co-accession digital records with associated paper materials; when to defer the accessioning of born-digital materials until after related paper materials or to simply accession portions of the born-digital materials as resources allow.

Additional considerations cover an understanding of both software and hardware that are available for capture in the repository, or which might be acquired to meet specific requirements of the records in question. This entails careful consideration of the costs associated with acquiring appropriate software and hardware, its effective testing and use, as well as understanding the types of media and transfers the institution will most often have to support. And finally, one must consider overall system capacity when anticipating a network-based transfer of data, rather than transfer by fixed media such as a hard drive.

Objectives

OBJECTIVE 1: Transfer records and gain administrative control

Outcome: *Records are transferred from the donor via electronic transfer or on physical media.*

Accessioning begins with the process of assuming custody of records and of relevant, related materials which have been identified during the collection development phase. In addition to data transfer, the archivist's role in this step includes the work done during the collection development stage to prepare for accessioning by determining that the transfer can be accommodated with the institution's technological infrastructure as well as verification that the transfer was completed.

A key difference between paper records and digital records is that gaining "physical custody" of digital records can include receipt of digital records on media as well as network- or web-based transfers of records from donors. Each type of transfer has different implications for actions that follow, and these will likely lead to distinct workflows that determine the sequence and type of subsequent accessioning tasks.[17] Accordingly, collecting institutions should be prepared to respond to any transfer scenario.

Additionally, archivists should recognize the distinction between digital materials and the media on which they are received. In other words, if an archivist receives an accession containing thirty floppy disks, those disks are

[17] For example, see "City of Vancouver Digital Archives System Workflow," (Fall 2010), http://artefactual.com/wiki/index.php?title=File:COV_Digital_Archives_System_Workflow_v1.pdf (last accessed July 13, 2011).

Evolution of Accessioning at University of Hull

Simon Wilson
Digital Archivist, University of Hull

Prior to involvement in the AIMS Project, Hull University Archives had no procedures or strategies for processing born-digital archives. In the early stages of the project, the Archives undertook a retrospective survey of collections to identify unaccessioned or unprocessed born-digital material. The first stage involved using CALM, the Archives' collections management system, to locate loose media among the already catalogued collections. Initially this process was a simple information gathering exercise with details about each media item (including the number and format of media) recorded in a simple Excel spreadsheet.

The spreadsheet revealed the range and type of media in the Archives' holdings, as well as unfamiliar file formats that needed further research. Archives staff found one 5.25" disc and, concerned about the long-term preservation and accessibility issues of the media format, identified a third-party vendor to retrieve the files from the disc. Staff then outlined procedures for removing the media from amongst its related paper files so that it could be stored separately in an appropriate environment. This "insertion sheet" provided information to both staff and users relating the disc and its content to the original location, as well as providing information on the accessibility of the content and procedures to request those files. (See sample insertion sheet in *Appendix F.4: University of Hull Insertion Sheet*).

This new workflow sought to identify key decision points and where information needed to be recorded. For example, we decided to document via a photograph each media item for curatorial and user needs, capturing the information on the item's label quickly and effectively – especially if faced with a pile of floppy disks. At this point the media are numbered in a simple running number associated with the accession number for that collection.

We endeavored to anticipate the information elements to include for each media type, but testing the process using actual media revealed the full picture and the realization that media formats vary widely. For example, Amstrad discs have three aspects to photograph (side A, side B, and the edge). A "clapperboard" template is used as a background in each photograph (see *Appendix F.3: University of Hull Digital Media Photography Form*), with the form field printed on a piece of transparency paper and annotated with a dry-wipe marker pen with each item's specific information, allowing the re-use of the form. Further testing then defined image quality parameters to ensure legibility of the labels without clogging up server space with unnecessarily large images. Once the documentation photo are captured, those files are named to reflect the media number and the shot perspective; these are then imported into the digital repository as supporting metadata. The entire workflow, processes, and forms are also clearly documented.

Although important, none of these processes actually tackle the contents of the media. The capacity to process a variety of media, including some legacy formats like 3.5" floppy disks and hard drives from PCs and laptops, was developed over time as the requirements evolved. The Archives were offered an old PC running Windows XP that was due to be decommissioned; with a built-in floppy and CD drives and USB ports, the mix of input/output options seemed ideal for reading some legacy formats. Colleagues in the University's Information and Communications Technology Department (ICT) assisted with cleaning the hard drive of old files that naturally accumulate through years of use and added an internal zip-drive to further increase the range of media the computer could handle.

To this "forensic workstation" we added software as well as several essential tools including FTK Imager (see Technical Evaluation and Use, *Appendix G*), DROID, and Karen's Directory Printer (see Technical Evaluation and Use, *Appendix G*, for reviews of several of these products). To protect the integrity of the data and reduce the possible impact of receiving material from third parties, the forensic workstation is a stand-alone machine. One of the main implications of this is the two-step process of downloading and then installing software or updates, like the DROID signatures, via a USB pen-drive. For each piece of software used, an in-house "idiots guide" is created to clarify the exact purpose of the software with-in the workflow, assisting with staff training and facilitating the assessment of other software.

the storage media, not the records; instead, the archivist must accession the records contained on those disks, in accordance with the collecting repository's collecting policy and the legal agreement.[18]

Decision points

In this early stage of the accessioning process, the archivist needs to determine what transfer methods are viable for the accession and make a determination of which to use. This is a continuation and finalization of the work begun in the last stage of collection development. Methods of transfer will be largely dependent on the nature of the material itself (is it on disks which can be physically transferred? is it cloud-based?), as well as the technological infrastructure of both parties. For example, large electronic transfers are infeasible over a low-bandwidth connection. It will be necessary for the institution to have at its disposal several different types of transfer methods in order to suit different scenarios as well as donor wishes.

Once the transfer takes place, where to store the material in the short-term (until it is ready for review) is another key decision. If the accession was of removable media, the archivist may choose to wait until they begin the next objective of stabilizing the records. If the transfer was electronic, the material may be downloaded to a quarantined computer (one not within the institution's network) until virus-scanning software can be run. It is wise to plot out the movement of data as it goes through the phases of accessioning (and indeed as it goes through the phases of arrangement and description and discovery and access) before beginning the transfer.

Tasks

- Confirm that documentation for the transfer of custody has been received and filed before moving forward with actual preparations for transfer.

  This would consist of a donor or sales agreement and supporting document, such as digital survey, enhanced curation elements, etc. Once this review is complete, staff are ready to transfer material in accordance with donor/transfer agreement(s) and accessioning policy.

- Determine how to transfer data.

  This will be highly dependent on the storage media used as well as the technological capacity of both the donor and institution. However, additional consideration should be made regarding the necessity of maintaining the security of transfer using encryption.

  Transfer methods may include direct transfer, transfer of physical media, or transfer of files onto transit media.

- If the collecting institution did not receive an inventory of the transferred records, create an inventory now.

  In the born-digital environment, the scale of transferred records may mean that establishment of this documentation must be undertaken using tools that can capture technical and administrative metadata automatically.[19] Tools such as the TAPER submission agreement

[18] This view does not preclude the possibility that the received physical media may have artifactual value or may otherwise be of curatorial interest. Additionally, depending on the donor or records creator, parts of the medium itself, such as the disk's label or sleeve, may be useful as a record itself. See Kirschenbaum, et.al. (2009, October) "Digital Materiality: Preserving access to computers as complete environment" for a more thorough exploration of these issues.

[19] In addition to the reviews found in *Appendix G* of this document, the *Practical E-Records Blog* contains reviews of several software options for this type of work: http://e-records.chrisprom.com/?page_id=175

builder[20] can also assist in systematically documenting accessioned records and their transfer terms.

- Verify that the transfer is complete and accurate using either a file manifest or checksums.

  The file manifest of the transferred files should contain technical details such as file size and dates created in order to verify that the contents of files were complete when compared against the originals. Checksums or hash values (a sequence of numbers generated by an algorithmic analysis of the data in a file) are probably the best indicator of the success of transfers since the checksums will only match if the data is exactly the same between two files.

- Document the success and/or failure of the transfer in the register, accession record, etc.

  Administrative control focuses on the documentation of the transfer. This can include the creation and maintenance of an accession file that documents the legal transfer, and the registration of the accession using a log book, database, or standardized form. The accession file should contain documents created in generating legal custody, such as the legal agreements and attachments.

OBJECTIVE 2: Stabilize transferred records

> ***Outcome:*** *Records have been prepared for long-term storage without any damage to the integrity of files, as evidenced through verification of checksum values. In addition, the institutional system is safe from any virus or malware that may have been part of the original transfer. System metadata is extracted and basic administrative control is applied through assigning identifiers and assessment of material.*

Upon or shortly following the actual transfer of physical custody, archivists need to establish physical control over the received records. This control needs to be understood as distinct from custody and primarily concerned with mitigating threats to preservation. The process of then stabilizing materials includes the safe extraction of records for long-term storage as well as the establishment of basic intellectual and administrative control. Throughout this process the integrity and authenticity of records must be ensured.

Physical control and stabilization includes assessing the condition of records and addressing issues identified during this initial condition assessment. Just as an assessment of preservation needs is critical when accessioning analog materials, a similar stabilization of threats is necessary with born-digital materials. However, unlike paper records, identification of potential preservation issues or threats for digital records can be significantly more difficult to ascertain. The condition of the physical media (if received) is only one dimension; other concerns include the presence of viruses and malware, or the presence of new or unknown file formats. While unknown file formats may not be malicious in intent, they may still prove problematic.

Similar to Roe's general statement that "[t]he archivist needs to avoid bringing preservation problems into an area where those problems may affect other records,"[21] archivists responsible for digital records must take additional steps to address these issues before proceeding with further work on the records. If these are left

[20] Tufts University Digital Collections and Archives, "TAPER: Tufts Accessioning Program for Electronic Records." http://sites.tufts.edu/dca/about-us/research-initiatives/taper-tufts-accessioning-program-for-electronic-records/ (accessed 9 August 2011).

[21] Roe, K. (2005). Arranging & describing archives & manuscripts. Chicago: The Society of American Archivists.

unaddressed, these issues can threaten the integrity of not only the newly received digital records, but also those already under the control of the collecting institution or within a given storage or preservation environment.

Decision points

As the archivist moves into the more technical aspects of the workflow and begins actually working with records, the need for a more sophisticated understanding of digital objects and the tools for working with them is required. If the transfer in the previous objective was of removable media, at this point the data should be transferred to production space, and checksums or hash values should be used to ensure this transfer is complete (see the discussion of checksums in the previous objective for further details).

The process of extracting records, particularly from older and obsolete storage media uses similar techniques to digital forensics.[22] Tools for digital forensics can therefore be useful to archivists. The transfer of data from media will probably include imaging or obtaining an exact, bit-by-bit replica of the original media. This so-called "forensic imaging" allows for the recovery of deleted files that have not yet been overwritten. This presents ethical problems for institutions when data is extracted beyond a donor's informed consent — particularly if the issue was not raised with the donor at the point of transfer, as might apply with legacy storage media.[23] As an alternative, "logical" copying, or simply copying the files that were part of the original agreement, is a better method of transfer in many situations.

Where the data is stored during this production phase is another decision, as it was in the prior objective. Institutional infrastructures may necessitate the creation of a separate production space for ease of access to material during accessioning. The length of time that data should be kept in this quarantined space will depend on institutional policies. For recently created data especially, this may need to be a quarantined space separate from the institution's network until virus-scanning software can be run — a crucial step in this objective. The handling of those viruses or malware should also be decided on ahead of time. Removing them is probably the most likely scenario, but it may be important to retain a copy of them in an inactive state as part of the Submission Information Package (SIP) for certain types of research activities (the history of software development, for example).

The handling of the physical media itself is another decision at this stage. Determining whether the media itself is significant is an open question.[24] Certain scholars may view the creator's annotation on labels or other aspects of materiality important. These may be captured sufficiently through digital photography of the media.

---

[22] Projects looking at forensic techniques within a specifically archival ethical context include:
FIDO at Kings College London: The Forensic Investigation of Digital Objects project aims to investigate the application of digital forensics within the working practices of a UK HE archive http://fido.cerch.kcl.ac.uk/
The BitCurator project: a joint effort led by the School of Information and Library Science at the University of North Carolina, Chapel Hill (SILS) and the Maryland Institute for Technology in the Humanities (MITH) to develop a system for collecting professionals that incorporates the functionality of many digital forensics tools. http://bitcurator.net/aboutbc/
Curator's Workbench, developed at UNC Libraries: software for capture and arrangement of submissions to a repository: http://www.lib.unc.edu/blogs/cdr/index.php/about-the-curators-workbench/

[23] For a full discussion of these issues, see Matthew Kirschenbaum, Richard Ovenden, and Gabriela Redwine, *Digital Forensics and Born-Digital Content in Cultural Heritage Collections* (Washington: Council on Library and Information Resources, December 2010),

[24] See Kirschenbaum, et.al. (2009, October) "Digital Materiality: Preserving access to computers as complete environment" for a more thorough exploration of these issues.

Project Xanadu: Loss and Recovery

Henry Lowood
Curator, History of Science & Technology Collections: Film & Media Collections, Stanford University

"What we're actually building at this point is only a part of Ted's original conception, though it's designed to be the first stepping stone to the whole thing."

– Chris Hibbert, post to comp.multimedia newsgroup, 30 March 1992 (from file on XOR hard drive).

Ted Nelson's Project Xanadu provided the original vision of hypertext as a system for document management, publication, linking and citation. Begun in 1960, the project to build Xanadu continued well into the 1990s. From 1989 to 1992, Autodesk funded Nelson's Xanadu Operating Company (XOC) to complete software development. However, when a new group of programmers primarily from Xerox's Palo Alto Research Center (PARC) joined the group in 1991, they abandoned the earlier version of Xanadu written largely by Roger Gregory and began a new version rewritten from scratch in PARC's new programming language, Smalltalk. This forking of the project eventually led to the collapse of the Autodesk-funded effort. Keith Henson, an XOC investor, encouraged the Palo Alto-startup, Memex to pick up the project In 1994. Memex licensed Xanadu from XOC and brought the Xanada project to its office space on California Avenue. Before long, however, the arrangement collapsed. The team disintegrated, with Nelson and Gregory regaining control of Xanada, which would finally be released as the open-source Udanax system in 1999.

Shortly thereafter, in 2001, the Stanford Libraries acquired the papers of Keith Henson and his wife Arel Lucas. The papers provided some documents from the history of XOC and included six hard drives, identified only as being from XOC in the mid-1990s. These mysterious hard drives were included in the AIMS project, because their source suggested that they might be significant for documenting the history of Xanadu. Moreover, the task of recovering data could well provide an interesting challenge. Indeed, the Stanford team was able to successfully image only two of the drives. Mechanical or formatting issues with the other four drives prevented access to the files on them. In order to learn if it would be possible to recover data from these drives, one was sent to Recovery Services, Inc. (RSI), which has a proven reputation in the area of data recovery services. RSI determined that we were dealing with "severe physical failures, some of them associated with read and write head errors." They concluded, however, that "it may still be possible" to recover at least some of the data. As stewards of the Henson Papers, we decided to cover the not insignificant cost (nearly $10,000) of the recovery option with RSI. We note that the expense of commercial data recovery may provide an obstacle for frequent use of this method.

The RSI effort was generally successful. It yielded three disk images, as well as capturing a significant number of files from the three drives from which RSI was unable to capture a complete disk image. Information gleaned from the recovered data reveals much about the provenance and significance of these hard drives. For example, several files document use of XOC's backup system and from file creation dates we learned that nearly all of the files were created between 1989 and 1993. Some files include header tags such as "historical" or "xanadu archive," so that we know that they were identified as being of historical interest. Specifically, many files contain source code and libraries in Smalltalk, with author names that correspond to the names of the XOC programming team during the Autodesk period. These files contain source code for the Xanadu version known today as Udanax Gold (formerly Xanadu 92.1), the version that was shut down when the Memex-based team disbanded. In addition to source code, the drives contain documentation about XOC, such as versions of a business plan that appears to have been written in 1982, text files from the 1983 "ninth printing" of Nelson's self-published (and increasingly rare) Computer Lib, draft chapters of Computer Lib and Dream Machines from Sept. 1984, and later documents such as descriptions of Xanadu and the work at XOC during the early 1990s. These documents will fill gaps in the historical record of XOR and the development of the Xanadu system and thus contribute to history of hypertext and related technology such as the World Wide Web.

In some cases, the destruction of the media may be a requirement. For example, if certain files were copied from a hard drive, but the rest of the drive contained sensitive information that would normally be removed and destroyed, the physical destruction of the media may be the only way to truly remove that data.

Finally, decisions about metadata and identifiers must be made. The level of metadata created during accessioning is going to be the one of the largest determinants of how much work is involved. The extraction of technical metadata can be achieved through fairly routine processes (see *Appendix G* for reviews of tools to do this extraction), but the institution may not need to retain all of the metadata that could potentially be created. Next, the archivist will assign identifiers, or a unique key for each an item in a repository; these identifiers may correspond with an existing scheme if one is available, but they may also need to be assigned separately during this process. The application of identifiers could be done at different aggregations, rather than at the individual file level (an identifier for the disk image, perhaps, with individual files referred to by their filename). The identifiers may be recorded in a media log or inventory, or they may be used to populate a basic finding aid or collection guide. The application of identifiers can be time-consuming, so staff resources should be part of the decision.

Tasks

- Remove media for separate accessioning workflow. This may represent a divergence with hybrid collections where media might not be discovered until later requiring retrospective accessioning.
- Photograph media to retain a record of any significant information found on carriers. This can also be useful in documentation.
- Assess the physical condition of material. Record any physical damage to storage media which may cause incomplete transfer/capture of content
- "Rehouse" material to ensure the physical and digital stability. Once the data is safely transferred to institution-controlled space, the archivist can begin the process of managing it:
    - Imaging or file transfer from received media
    - Bit-level (forensic) disk imaging
    - Filesystem-level (logical) disk imaging
    - Direct copying
    - Verify transfer through use of checksum validation
    - Physically rebox or rehouse media if it is to be retained; destroy if required

    Media may be destroyed if necessary through wiping, overwriting, or physical destruction.
- Stabilize the data by running virus and malware checks and removing these materials as appropriate.
- Identify files in obsolete or unknown formats for future normalization or migration.
- Harvest metadata from files and file system. Create a record of the following:
    - High-level inventory of filenames
    - Timestamps
    - Technical metadata
    - Filesystem structure
    - Checksums/hash values for the media (in addition to values for each file)

- Repeat any failed processes, if possible.

OBJECTIVE 3: Intellectual control and documentation to support further processes

> ***Outcome:*** *Actions taken and issues to be addressed in future processing are documented in a standardized format. Prioritization of next actions needed, especially appraisal, arrangement, and description, are clearly indicated for planning purposes.*

During accessioning, archivists have the opportunity to perform an initial assessment of the content of the records. This basic assessment allows for the possibility of creating high-level description for the accessioning (including identification of the creator) and for estimating the extent and dates of creation, the intellectual property status, and an overview of contents of the accession (such as the types of records it contains). In addition to assisting administrative control, creation of a detailed inventory establishes a basic level of intellectual control over transferred records. However, depending on the original computing environment in which the records were created, the archivist may have significant difficulty creating an inventory even at the most basic level of a listing of directories and files.

Establishing physical, administrative, or intellectual control over digital records may require archivists to undertake processes distinct from or otherwise inapplicable to other formats of records. While these processes may assist with maintaining the authenticity, reliability, or viability of the records, documentation of these actions is also essential.[25]

Maintaining the authenticity and reliability of the records is particularly more complicated in a digital environment. The nature of digital material is in itself not static. Digital objects are more than just a static series of bits, but instead a dynamic interaction of data, system, and software. Each time a digital object is viewed, it is in essence a re-creation of that complex interaction of variables. As such, if one is not careful, technical characteristics of the data can be altered when a file is viewed. In addition, the higher level of fragility of digital data, due to threats of bit corruption and obsolescence, equate to a higher level of risk of data loss. To mitigate this risk, archivists must perform integrity checks on a regular basis throughout the digital lifecycle. It is therefore essential to document technical characteristics of data at accessioning so that these future checks can be verified against accurate original data.

Decision points

When entering the intellectual control stage, the crucial issues are what type of information should be documented and what format should that documentation take? Documentation that is already in use at the institution may be appropriate in some cases, but this is unlikely to record the correct type of detail. Records could include written documentation or reports, spreadsheets, forms, or other types of machine actionable documentation such as metadata extracted by a software tool in the previous objective and reported in an XML format. Typical data that could be tracked include:

---

[25] See Matthew Kirschenbaum, Richard Ovenden, and Gabriela Redwine, *Digital Forensics and Born-Digital Content in Cultural Heritage Collections* (Washington: Council on Library and Information Resources, December 2010), p. 38-39.

- Files –
  - Listing of files, summary of labels, or categories
  - File formats
  - Structure of transferred materials
  - File system/directory hierarchy

- Physical Media –
  - Photograph of media
  - Inventory
  - Media log
  - Separation sheet

The other crucial decision at this stage is how the documentation will be used. Records created as some sort of structured data (XML or spreadsheets) could be ingested into part of an archival data management repository. On the other hand, they could simply be filed for future use when processing. The documentation could also be used differently in the future. For example, it could be used to triage or prioritize processing needs or to determine that future processing is unnecessary. Understanding the ultimate usage of the records will help to inform the creation of them.

Tasks

- Create accession records documenting transferred materials, both in terms of files and physical media.
- Document potential restrictions or material that was accidentally taken in, either through physical transfer or disk imaging, and whether donor needs to be contacted.
- Identify duplicate assets (using batch processes if possible) using a method such as a checksum.
- Create an audit trail of actions performed during accessioning; include actions that fail.

  Documentation of processes undertaken during accessioning provides context to support decision-making during future processes. The level and method of documentation will be somewhat determined by institutional practices, but new types of documentation may need to be developed when beginning to work with born-digital materials. Documentation can also serve to inform the donor and researchers of any records that were not successfully transferred or that had to be removed.
- Document needs for future processing, and if possible how they may be addressed, including:
  - Arrangement and description
  - Appraisal
  - Discovery and access

  Many archivists will find it impractical to address all of the needs of a given accession during accessioning, and many collecting institutions will need to prioritize work required across their collections. Accordingly, archivists should identify the appraisal, arrangement, description, and preservation needs of the accession, and, if appropriate, its associated collection, and document those needs in a systematic way. An example of this documentation can include, or alternately inform, the development of processing plans (see *Appendix E* for examples). Archivists responsible for accessioning digital records should consider performing these assessments collaboratively with colleagues responsible for collection development, arrangement and description, reference, preservation, and information technology as appropriate.[26]

[26] Roe, p. 55-56.

- Send donor an acknowledgement of successful transfer.

  Maintaining the institution's relationship and trust with the donor by letting them know that the transfer of records was successful is encouraged. Since the receipt of digital materials is also not a physical process, and therefore may not be obvious, it is important to document that the process has been completed and was successful.

OBJECTIVE 4: Maintain accessioned records

> ***Outcome:*** *Records are safe and secure in stable medium- and long-term storage. They also remain viable and accessible for further work.*

To successfully conclude the accessioning stage, data must be stored in a stable environment and a regular routine of maintenance activities begun (systematic integrity checks, for example). These processes ensure further access and viability of records and assets. A key part of this process is the storage of accessioned records and assets in a stable environment. Ideally, this type of routine is managed by some type of preservation or maintenance repository. However, an institution may also place copies of accessioned data in a separate medium-term storage environment or production space to await further processing. Wherever the material is stored, the repository should record the storage location, the success of the transfer to that location, and any transformations of data undertaken for preservation purposes. These maintenance activities and others are incorporated into the Archival Storage and Data Management functional areas of the OAIS model.[27]

Decision points

Decisions about the proper storage for maintenance may not necessarily be made by the archivist alone or on an individual collection basis. Instead the institution may have or may be planning a preservation repository for storage. The questions for the archivist are whether or not it will it be adequate for these materials and how will material be transferred to it — or from it — as needed.

Normalization of material may also be addressed at this point. During the stabilization stage, material that was in need of normalization may have been identified. At this point, the archivist must decide if normalization will be carried out before transfer to long-term storage. Although some types of normalization will be a matter of established workflow or policy, when new formats are encountered some reconsideration may be necessary to determine how normalization should take place.

Tasks

- Perform necessary normalizations to preservation and access formats.
- Create the Submission Information Package (SIP) that will be stored for future processing. This will include the accessioned data as well as metadata created during accessioning.
- Transfer the package to medium- and/or long-term storage environment.

---

[27] See Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-B-1, January 2002, sections 4.1.1.3 and 4.1.1.4. http://public.ccsds.org/publications/archive/650x0b1.PDF (Accessed 20 June 2011). For a more detailed discussion of maintenance activities, see Fedora and the Preservation of University Records Project. 3.1 Maintain Guide, Version 1.0, 2006. http://hdl.handle.net/10427/1286 (Accessed 29 July 2011).

Where the SIP goes at this point will be a matter of local infrastructure. If the long-term repository has adequate tools to allow for the downloading of this material for future work in arrangement and description, then long-term storage may be used. Alternately, the institution may wish to store the accessioned material in a medium-term space, or in the same production space that was used for accessioning until further work is done.

- Verify success of transfer.

  The use of checksums of file manifests should again be used for this verification.

- Record the storage location, any normalization, and the success of transfer in appropriate metadata records.

  This metadata may be within the institution's preservation repository, within an archival data management system, within accessioning records, within finding aids or other collections, and/or in other institution-specific systems.

3. Arrangement and Description

Definition and Scope

Arrangement and description: the process undertaken by an institution to establish intellectual control of the material following the physical control secured during accessioning. It also prepares the material for discovery by providing the user with information and context about the records, and prepares for access by applying appropriate restrictions. All of this must take into account the broader policy and technical infrastructure.

Preface

Although the processing of born-digital material requires new skills and technologies, the basic practices can still be addressed within a traditional processing workflow. The crucial differences between traditional paper and born-digital processes stem from the wide range of file types, the sheer volume of born-digital material, and the difficulties inherent in viewing the contents. Once archivists determine which tools to use and how to integrate them and modify them if necessary, the key challenge becomes balancing the needs of the material with available resources. There are tradeoffs or compromises to be made as not all institutions will have the necessary capacity or infrastructure to cope with the expected deluge of digital content.

At its most fundamental, arrangement in the digital world is the representation of relationships between items. The organization of material into a "folder" and "file" is representational only — the data of the digital items themselves are not organized this way on the physical hard disk or other storage medium. Metadata captured at the point of accessioning can be reused during processing to represent this organization. Born-digital material can have multiple arrangements (or rather, multiple arrangements can be represented), such as the original order of the files as they were received or a different order applied by the archivist. Files could even be re-organized (or differently represented) by the user through manipulation of the metadata and data online — for example by sorting a collection into date order, by title or by file format, etc. The AIMS project partners decided to use the term "intellectual arrangement" in describing the work of the archivist (the activity and its result) to emphasize the fact that the records themselves are not manipulated.

It was beyond the scope of the AIMS project to consider specific cataloging or description standards for born-digital material. This was partly an acknowledgment that cataloging standards such as DACS and ISAD(G) are intended to be format agnostic, and also a recognition that this aspect will develop and evolve through local practice and conventions.[28]

[28] For example see the *Workbook on Digital Private Papers* produced by the Paradigm project, http://www.paradigm.ac.uk/workbook/cataloguing/index.html accessed 5 July 2011.

Keys to Success

The most crucial factor for the success of this function is the full implementation of processes for managing born-digital material in collection development and accessioning. If records cannot be captured and stabilized, they cannot be analyzed and processed. Equally important are the many other steps and resulting forms of documentation that enable the institution to maintain the integrity and authenticity of transferred records.

Success within arrangement and description of born-digital material can be described in the same way as traditional archival records:

- *Preserving the context in which records were created, managed, assembled, or accumulated irrespective of the format of the material:*[29] The preceding processes in the AIMS framework focused on gathering the evidence of this context and ensuring that the metadata embedded within the files is not lost or altered as the material is transferred to the institutional storage environment. With hybrid collections there is the added complication that the physical context for paper and born-digital material is likely to be different.[30] As well as preserving the context(s) of paper-based and born-digital material, it may also be necessary to gather evidence of the relationship between the two.
- *Establishing intellectual control over the material:* The understanding of the collection developed during the Accessioning stage is deepened and solidified during processing. The work of establishing intellectual control of records includes assessment and identification of the types of material in the collection, appraisal of the records' value and relationships within the context of the collection, arrangement of these materials in the best way to preserve context while providing organization and access, and finally documenting decisions made and knowledge gained about the collection.
- *Provide a finding aid or other means of discovery:* To ensure that born-digital collections are accessible, some means of discovery must be provided. While this is traditionally provided by means of a finding aid, which is also the product of establishing intellectual control, this does not mean that access cannot be granted before intellectual control is established, nor that the full extent of information gained during intellectual control be used for discovery. However, it does follow that tools and processes for arrangement and description need to take account of discovery routes. There are three relevant questions here: where descriptive metadata will reside, how it will be accessed (in an EAD guide? in a catalog record? attached to the digital files?), and whether users will be enabled or permitted to search the content itself (if it is textual).

In addition to these activities, success will also be contingent on having appropriate guidelines, established before beginning the work of arrangement and description. These guidelines should be developed with reference to institutional policies and to the curatorial areas or record creators for which the institution is responsible.

---

[29] Roe, *Arranging and Describing Archives and Manuscripts* (2005): 56.

[30] Heather MacNeil, "Archival Theory and Practice: Between Two Paradigms," Archivaria 37 (Spring 1994): 10. See also Zhang, Jane. *The Principle of Original Order & the Organization and Representation of Digital Archives*. Dissertation, Simmons, 2010: 178. In her research investigating the value and application of the principle of original order to born-digital materials in file directory systems (in personal record-keeping environments) and file classification systems (in organizational environments), Zhang observes how the file directory system functions as a virtual counterpart to the traditional paper filing system.

Technical Development: Functional Requirements for Arrangement and Description

Early in the project, the AIMS partners recognized that there were few options to facilitate the archivist's task of comprehensive intellectual arrangement and description of born-digital archives. The partner team acknowledged that other digital archives projects would greatly benefit from tool development in this area as well as the potential for a Hydra-based solution, and therefore decided to assign a considerable amount of project time to develop functional requirements for an arrangement and description tool.

Though three of the four AIMS partners (University of Hull, Stanford University, and University of Virginia) are also the three key institutional members in the Hydra partnership, the functional requirements for the arrangement and description tool were ultimately created to be agnostic in terms of technology. This allowed the digital archivists and other AIMS participants to be more flexible and focus on what was important or necessary without getting distracted by perceived limitations within Hydra, the Fedora digital repository environment on which the Hydra framework is based, or other software in use or familiar to the AIMS partners (such as Archivists' Toolkit, CALM, Curators' Workbench, or Forensic Toolkit). By focusing on the requirements and principles of archival practice and emphasizing commonality within the archival profession, the team endeavored to think broadly about what archivists outside the AIMS partnership would need to support the arrangement and description of born-digital archives. The requirements were thus written to describe symptomatic needs of ongoing work, and written whenever possible to reflect individual tasks to be completed. The writers of the functional requirements were asked to supplement these tasks with lists of preconditions, including required inputs and measurable results, including required outputs. While the team intended to be technologically agnostic, tasks could also be supplemented to provide examples of how existing software supported a particular task. The writers of the functional requirements found this particularly useful when trying to provide examples of interaction paradigms that would be familiar to other archivists.

Coordinated work on the functional requirements began at the AIMS partner meeting in September 2010, which was preceded by some background work that included functional decomposition of arrangement and description workflow. Most of the work was coordinated online and was supplemented by conference calls. A partial in-person meeting at Stanford University after the 2010 Digital Library Federation Fall Forum led to the identification of the following set of high-level requirements:

- Graphical User Interface[31]
- Viewing of technical metadata
- Viewing and editing of descriptive metadata
- Management of access rights and restrictions (by creating or editing administrative metadata)
- Viewing files or a representation of them
- Exporting metadata (for example EAD)
- Importing metadata (for example EAD)
- Creating reports (for example relating to file formats, dates or restrictions)
- Viewing email to enable processing
- Identifying duplicate files
- Viewing application metadata from files (e.g. filenames assigned to titles)
- Creation of new objects[32]

[31] to show a representation of the original arrangement of files and the new intellectual arrangement, as created

[32] This functional requirement does carry some implications for (or assumptions about) the environment in which the digital material is stored during arrangement and description and in the longer term. Fedora and other digital repositories enable files to be grouped in hierarchies. In order to define and describe the groups (for example series and sub-series in archival terms) new digital objects need to be created to contain the metadata.

(Technical Development: Functional Requirements for Arrangement and Description - continued from previous page)

Following this meeting, the functional requirements were collaboratively written using Google Docs over a period of several months. The digital archivists were primarily responsible for this work, but the functional requirements were also reviewed and updated by other AIMS participants, as well as by non-AIMS colleagues at the partner institutions. Following this collaborative effort, the digital archivists edited the document and generated a prioritized list of high-level and lower-level requirements, the result of a survey of AIMS staff. Ultimately, the digital archivists presented these prioritized functional requirements at the Hydra partner meeting hosted by University of Virginia in February 2011. The following represents the final prioritized list:

- Assumptions
  - Graphical user interface
- Essential Functionality (tool cannot function without these)
  - Presentation and manipulation of Intellectual arrangement
  - Viewing and editing of descriptive metadata
  - Allocation of actionable rights and permissions
- Important Functionality
  - Appraisal, for example by tagging files to be removed and activating batch delete
  - Viewing of technical metadata
  - Viewing of files or representations (with qualifications: focus on providing viewing for browser-renderable formats and an extensible framework to add other viewers later)
- OK / Depending on Resources
  - Creating reports
  - Importing metadata (other than EAD - mostly for entity extraction)
  - Searching within files
  - Batch application of metadata from Files
- Lowest Priority
  - Importing (not EAD) (other than entity extraction)
  - Viewing of e-mails within context (or other record formats such as databases)
  - Exporting (other than EAD)

The full functional requirements document can be found in *Appendix H.1*. Its overall structure is best understood by recognizing five key areas within each section of the document. First, there is the *overview*, which describes the functional area within a given section. This section provides context for the *tasks* that follow, which define tangible types of activity within a functional area. Some tasks contain *user stories*, which describe a hypothetical user needing to accomplish a given task, as well as expected application behavior. *Screenshots* demonstrate some aspects of the task using existing software. To provide further background regarding some for the decision made by the team, *questions and comments* by the AIMS participants can be found throughout the document. The functional requirements should be viewed as a critical output of the AIMS project, existing to inform not only the development of Hypatia going forward but other applications as well.

Appraisal and analysis tools

Appraisal and analysis of files during the arrangement and description process is critical both for identifying formats for preservation and for identifying restrictions required to ensure appropriate discovery and access. The content of individual files may be appraised or analyzed with a file viewer or appropriate software, depending upon whether disk images have been created during accessioning or whether files have been normalized into a standard format and replicated. There are commercially available forensic tools with built-in file viewers and stand-alone file viewers that may be used exclusively for this purpose.

The AIMS framework was purposefully developed to be software-agnostic in order to be as generalizable as possible. There are various tools that can be used to process hybrid collections and discussion in this section is supplemented by evaluations of specific tools available for use in appraisal, arrangement and description in *Appendix G*. Approaches include a traditional authoring tool for purposes of arrangement and description supplemented with external tools for appraisal/analysis of born-digital material. This strategy is viable for reasonably small numbers of media and files, but it is not scalable.

Another option is an archival data management tool, such as the Archivists' Toolkit (AT) or CALM. Once again, this strategy is viable for reasonably small numbers of media and files, but is not easily scalable.[33] A third option that has been explored at Stanford and at other institutions, such as the British Library, is to use Forensic Toolkit, a commercial forensic analysis software package, for purposes of archival arrangement and description. This strategy also has limitations (see *Appendix G: Technical Evaluation and Use*).

Other prototype tools have been developed for specific aspects of the workflow, but few have tackled arrangement and description. This gap prompted the AIMS team, lead by the Digital Archivists, to draft functional requirements for a tool that would enable archivists to arrange and describe born-digital materials natively in the Hydra repository environment. The functional requirements specified include:

- to import any existing description in EAD
- to add metadata; to set rights and restrictions
- to represent and manipulate directory structures and descriptive metadata, much as one can in a Windows environment, with drag-and-drop and other features
- to export EAD.

The value of these features becomes apparent when one considers the labor required to work with large volumes of files and to integrate born-digital and analog material in hybrid collections.

Initial development of a Hydra Head called Hypatia to meet these requirements has begun as part of the AIMS project[34] and an overview of work completed to date is included (see *Appendix H.3*). A summary narrative overview of the development of these functional requirements is found in "Functional Requirements for Arrangement and Description" on pgs. 31-32 and the full requirements themselves are included in *Appendix H. 1*. The development of Hypatia will continue as an element within the Hydra project, with the continuing involvement of some members of the AIMS team as advisors, reviewers, and testers from an archival point of view.

Objectives

OBJECTIVE 1: Prepare for processing

> ***Outcome:*** *Files and their technical metadata, acquired from the accessioning stage, can be viewed by the archivist and descriptive metadata can be created and/or edited.*

This preparatory or pre-processing stage includes the retrieval of collection material and all of the supporting documentation generated in the preceding stages. The process is equally applicable to paper and born-digital material and is critical to begin planning for arrangement and description.

---

[33] ArchivesSpace, the grant-funded Archon/AT merger, has draft specifications for digital objects at http://archivesspace.org/documents/specifications (accessed 30 November 2011). Colleagues at Hull University have contributed to a CALM digital records working group looking at aspects of integration of CALM with a digital repository. Although this work is ongoing, it has already led to the development of an API for data exchange.

[34] See the Hypatia wiki at https://wiki.duraspace.org/display/HYPAT/Home and the Hydra site at http://projecthydra.org/

Arrangement and Description Case Study: The Papers of Stephen Gallagher

Simon Wilson
Digital Archivist, University of Hull

Stephen Gallagher is a novelist, screenwriter, and a University of Hull alumnus. He deposited his paper archives with Hull University Archives in 2005, and in 2010 he donated born-digital material (14,320 files, 13.6GB) that was deposited via an external hard drive.

Although much of the born-digital material was comparatively well ordered, his more recent work was stored its own distinct folder, reflecting his frequent consultation of these files. In discussing his approach and methodology, it was clear that each of his works was seen as a distinct "project" and that there were often multiple projects at different stages of development at any one time. For example, a short story (that was subsequently dramatised for radio and then also for TV) would represent three separate projects.

In devising an intellectual arrangement, the Archives sought to create a framework that was logical, that could accommodate future accruals, and that would help researchers to locate the material they wanted. We proposed that the first level of arrangement (sub-collection) should reflect the nature of the output - whether a short story, novel, radio or screenplay. Each project then formed a discrete archival series within the appropriate sub-collection. Rather than describe each born-digital item, the collection guide includes a description of the short story or novel and just an outline of the range of born-digital material in that series. Nonetheless, each individual file had to be examined to ascertain the file's actual contents and to ensure it did not contain potentially sensitive material.

We faced two technical challenges – over 300 files created using specialized screen-writing software (FinalDraft) and 39 Amstrad disks. After consulting the donor, we were confident that although these represented a problem for long-term preservation, neither would impact the arrangement or description of the collection.

Other issues encountered focused on copyright: first, the files included some material from third-parties. Additionally, the donor would recycle ideas between projects — an idea that was unsuccessful in one guise could re-appear several years later. We discussed an appropriate time gap between creation and releasing the material online so that we would not impinging on the donor's intellectual property rights. [35]

As with paper archives, each collection is unique, and we found this arrangement and level of description were appropriate for this collection. Another lesson learned: having the same personnel arrange and describe the paper and digital components of hybrid collections made integration much easier, as they drew from their familiarity with the content and the creator's working practices to successfully process the collection.

Decision points

To complete the pre-processing stage, the archivist must determine what tools are needed. In the born-digital environment, it may be difficult to even view files and their associated technical metadata. If any preservation activities such as migration or normalization need to happen at this stage (if they weren't already undertaken during accessioning), specialized tools may be needed. It is likely that each institution acquiring born-digital material from a range of sources will find itself with media that it cannot handle or files it cannot read. Although this situation is not unique to the born-digital environment (analog equivalents include material in foreign languages), it is an aspect that needs to be addressed in terms of donor expectations and users access to the collection; the institution must consider what is a reasonable level of effort or expense to attempt to rectify the problem.

As has already been discussed, tools exist to make these tasks possible, although limitations

[35] Making previously unpublished material freely and universally available on-line is regarded in the UK at least as constituting publication. Padfield, Tim. Copyright for Archivists and Records Managers, Third Edition, London 2007, pp93-96

due to unusual or obsolete formats may still need to be addressed. Depending upon the context of the unreadable material in a particular collection, the decision might be made to continue processing the remainder of the collection or to stop until this issue is resolved. One of the paradoxes of born-digital material is the additional technical effort required to convert or migrate content to a readable format, only to then possibly make the professional decision that the item is not wanted. There are many parallels in the analog world, especially among audio-visual material.

Tasks

- Retrieve the material, which may require placing access restrictions on the material until processing has been completed. A media log or similar documentation will help to ensure that no materials still on legacy storage media are overlooked.[36]
- Review supporting documentation and metadata generated during the collection development and accessioning stages. This may include:
  - existing information and structure for related paper material already held by the institution
  - photographs of the storage media
  - file manifests and file-type or file-format analysis
- This information will direct future tasks and decisions, including which tools are required and whether any of the files need to be transformed or migrated so that they can be viewed — critical for description and before any appraisal decisions can be made.

OBJECTIVE 2: Plan for processing in accordance with policy and technical framework

> ***Outcome:*** *Documentation that will guide the processing of materials is produced. This documentation may include a survey of the collection (documenting the context, structure, content, and condition of material), a processing plan (documenting the recommended arrangement, description, and appraisal where applicable), and the rationale for the recommendation. It may also identify work which is beyond the bounds of current capabilities, for technical or other reasons.*[37]

The planning stage is at the center of processing. Decisions made during this planning stage will determine the work done in subsequent steps. In fact, most tasks at this stage are related to gathering information for making these decisions. The planning stage is a good time to test out new tools for working with born-digital materials to determine what will work best for the institution and its workflow. At the close of the planning stage, the archivist should be able to seamlessly move on to implementing the suggested arrangement and creating the descriptions at the proposed level.

[36] It is expected that in many institutions the largely technical work of accessioning may be undertaken by different staff to those involved in the arrangement and description of the material.

[37] This is equivalent to analog materials which cannot be processed until conservation measures have been carried out.

Decision Points

As with paper and other records, the archivist's key decision will be the overall processing or cataloging strategy which will determine the level of arrangement and description and appraisal. As each collection is different, the strategy is more akin to a series of principles for the archivist to consider. These might include the level of cataloging effort required, the extent of integration with previous accessions within the same collection (if applicable), and whether to retain both paper and born-digital versions of the same item. Other strategic decisions include how to reference the existence or location of born-digital content within the finding aid or other means of discovery, and how to reference individual files or series of files within preservation environment where they will be stored.

An explicit strategy for identifying, determining and applying access restrictions is also paramount in the born-digital realm. The likelihood of the material containing sensitive information should first arise during consultation with the donor. Experience with paper-based records reveals that the donor is not always aware of the exact nature of the content being transferred; the sheer volume of born-digital material exacerbates this situation. Although the application of access restrictions will not be determined by its format, born-digital material does offer new opportunities for automatically detecting the presence of potentially sensitive information. Tools such as Forensic Toolkit (see technical reviews in *Appendix G*) and EnCase Forensic offer the ability to conduct pattern searches for things like social security or credit card numbers and keyword searches, including searching on related or fuzzy terms.

For most institutions the sheer volume of material received will make the manual checking of each file unmanageable. In such situations the institution may wish to adopt a risk-management approach, weighing the risk of sensitive material being discovered against the cost of manually checking each file or introducing checking as part of the process to provide access to the files.

As born-digital material becomes more common, and supporting documentation becomes more comprehensive, the appraisal/analysis strategy becomes an important decision point at this planning stage and documentation of restrictions becomes an important outcome. Institutional processing guidelines will need to be sufficiently flexible to respect the various kinds of restrictions that may exist for this material. Each collection will continue to be assessed on its own merits, but, as born-digital collections become more common, the body of work and evidence that the archivists can draw upon will grow.

Tasks

- Review relevant policy, guidelines, and supporting documentation created during collection development and accessioning.

  The workflow for hybrid and born-digital material is more complex than for paper-based collections, placing greater emphasis on the documentation generated during collection development and accessioning processes and the underlying policies. The following policies may be applicable:

  - digital preservation policy
  - processing guidelines
  - integrity/authenticity criteria

- copyright legislation and institutional restrictions on content.

The preservation policy is relevant here because, upon closer analysis of files and associated technical metadata, formats not supported by the institution and/or digital preservation repository may be discovered. The decision must be made whether to extend support to this previously unsupported format or to discuss with the donor possible alternatives.

If significant time has passed between acquisition and processing, or if a member of staff is appraising/analyzing the born-digital material for the first time, all of the supporting documents created through collection development and accessioning processes will provide important context for planning. The media log and record of actions taken document the custodial history for media in the collection and provide the basis of support (or lack of support) for recording the integrity/authenticity of files at the aggregate level in the finding aid. It is also important to gather information provided by the donor and recorded during collection development, or, if this was not the case, to approach the donor retrospectively (if possible).

- Set strategy for determining and applying restrictions.

  The needs of born-digital archives appear to be at odds with minimal processing trends for modern paper-based collections. This is most evident with regard to the processes of appraisal and identifying and setting restrictions. Restrictions may be defined by file format (set by digital preservation policy) or content restrictions (set by legislation and/or the institution). Institutional collecting guidelines and donor/purchase agreements may also set broad restrictions on digital content by date of content and/or class of material. Given the quantity of files to be appraised and analyzed, the development of criteria, tools, and automated processes which enable this work to be done in bulk is key to the application of an MPLP strategy.

  Collecting guidelines and donor/purchase agreements may also identify content to be de-accessioned when it is found following transfer/capture. Additionally, the need for restrictions on retaining or providing access to certain kinds of content in disk images will be identified at this stage. If restrictions are set against access to deleted files recovered within disk images, for example, those deleted files will need to be identified at this stage if they have not yet been filtered out and omitted from the arrangement and description.

- Assess the born-digital records and relationships with other material and previous accessions (if they exist).

  The archivist's ability to determine the context of born-digital material will depend in part on the collection development approach. Context and structure will be easier to determine if records have been acquired through a snap-shot accession or retired computer, supported by site visits and communication with a donor, possibly documented by a records survey, directory lists, or email. Accruals of born-digital archives may be more difficult to deal with — for example successive, iterative 'snap-shot' captures of active computers, drives or servers. However, as archivists we are used to acquiring material with little evidential context and we must expect this in the digital age.

- Assess the integrity/authenticity of the records.

  The integrity/authenticity of records will have formed a central element of the negotiation with the donor during the collection development process and is likely to be reconsidered during arrangement and description. Criteria have been proposed for assessing authenticity of electronic records in electronic records systems, but there has been limited discussion in the archival community and between members of the archival and scholarly communities on assessing the authenticity of born-

digital material in personal recordkeeping environments and in archival collections.[38] Further consideration is required in this area.

- Determine or propose an arrangement.

  Once a collection has been surveyed, an arrangement may be determined and proposed. There are numerous factors influencing arrangement strategies with born-digital material in hybrid collections, including the principles of provenance and original order, institutional collecting policies and processing guidelines, collection development approaches, and media formats. AIMS project partners support the broad Paradigm recommendation to respect context and content first in combining and arranging hybrid collections.[39]

  At first glance it would appear to be easier to respect context for born-digital material than it might be for paper, since the organization of the files on their carrier media would seem to imply an intentional arrangement. Discussions with the donor will identify whether this was in fact intentional or simply the result of other work practices. For instance, files may have been saved on whatever floppy disk was at hand, or disks may have been filled sequentially until full. In other cases, files may have been merged from other media carriers onto a new one without their original contextual arrangement.

  The situation becomes more complex when dealing with files from multiple sources, for example, network and personal files saved on separate servers within an organization or from current and back-up data sources. In addition, born-digital files must be considered along with their paper counterparts. As archivists we do have some experience in taking material from multiple sources — for example, complicated business archives that contain records from mergers and acquisitions. Archivists have a responsibility to consider how to arrange records from computers with other storage media and paper in the collection[40] and to accurately represent this context to researchers.

  As with paper collections, the approach adopted will vary depending upon the nature and volume of material under consideration. It remains to be seen if context becomes easier with larger volumes of material stored on external drives or more complicated as a wider range of sources are taken into account.

- Determine the level of description.

  The level of description should be heavily influenced by the policies and approaches that have evolved within each institution. The overall aims and objectives of this step remain unchanged by format. Factors to consider include the work necessary to create description and the value it will provide for research.[41] The level of description possible and desirable may vary between paper and born-digital material, reflecting the variations in available resources to undertake the descriptive work and also a belief that full text indexing of the born-digital material empowers the user while at the same time

---

[38] See InterPARES 1, Authenticity Task Force, "Requirements for Assessing and Maintaining Authenticity of Electronic Records" (March 2002). Forstrom, Michael. "Managing Electronic Records in Manuscript Collections: A Case Study from the Beinecke Rare Book and Manuscript Library," *The American Archivist* 72, 2 (2009): 460-477. And Kirschenbaum, Matthew, Richard Ovenden, and Gabriela Redwine. *Digital Forensics and Born-Digital Content in Cultural Heritage Institutions*. Washington, D.C.: Council on Library and Information Resources (December, 2010): 32-39.

[39] See the *Paradigm* project's Workbook at http://www.paradigm.ac.uk/workbook/

[40] Kirschenbaum, Matt, et al. *Digital Materiality: Preserving Access to Computers as Complete Environments.*

[41] At Hull University Archives, for example, the decision was made to extend a simple priorities scoring matrix for cataloging to include born-digital material and this has been added to their processing plan template.

reduces the need for the archivist's description. Researchers have examined the potential to repurpose file and item-level metadata in the recent Pedals project.[42]

OBJECTIVE 3: Process material and metadata in accordance with processing plan and policy and technical framework

Outcome: *A collection is processed with added descriptive and other associated metadata and documentation on record removal.*

At this point the archivist can begin implementing the processing plan created within the previous objective. Since much of the intellectual work of processing was completed in the planning stage, the work now is to implement the plan. However, the process is iterative and some specific issues (such as the handling of content to be restricted or removed) must be faced at this stage.

Decision Points

Determining what to do with records marked for removal during appraisal is a major consideration at this stage. The specific scope and characteristics of the removed material will not be fully known until after arrangement is carried out, so while some of the issues may be anticipated, the best solution or methodology will only become clear as the work progresses. Secure deletion of records may be undertaken through several methods and the archivist will have to determine the most efficient way to completely remove data from the server space or other storage media depending upon the agreement with the donor.

The timing of preservation and appraisal activity, including integrity and authenticity verification, is another issue to be considered during processing. The likely range of these activities may be anticipated during the planning phase, but the full extent of what actually needs to be done may not be clear until after an appraisal of the material has been completed. The timing of any file migration or normalization work will also be determined by the access policy for these particular records within this particular collection. A workflow could include migration of material to an access format immediately after the arrangement and description of the material, or this could take place at the point an access request is received. Both of these approaches have strengths and weaknesses and are very much dependent upon the availability of tools to undertake the work and the mechanism for delivery and access.

Tasks

- Identify records for retention, restriction, and removal (if possible); identify relevant levels of access and edit metadata accordingly.

  Any restriction or removal identified in the planning stage now needs to be carried out. The need for a functionality to apply access restrictions to material in born-digital collections was identified by the AIMS partners as one of the most critical aspects for the arrangement and description tool (see *Appendix H.1*). Restrictions may need to be defined by the dates of the born-digital material and/or on classes of user[43] and be capable of being read by humans and processed by machine.

[42] http://www.pedalspreservation.org/

[43] Classes of user might distinguish between the owner/creator of the material, archivists and authenticated or non-authenticated users. In the UK, the Freedom of Information Act generally prevents the provision of access to a specific body of material to some members of the public but not others.

Discussions with the donor in the collection development stage, and subsequent review of the content by the archivist, will lead to the identification of appropriate levels of access to the material within the collection. As with paper material there is a need to allow this to be set at varying degrees and to recognize that this may change over time – for example to comply with relevant legal restrictions that close material for a specified duration.

Four levels of access were proposed within the functional requirements for an arrangement and description tool: ***discover***, which would allow items to be identified by a search of metadata; ***view***, which would allow metadata to be viewed; ***render***, which would allow browser-renderable representations of content to be displayed (and would also permit searching of content alongside metadata if systems enable this); and ***download***, which would allow associated files to be downloaded.

- Apply copyright restrictions.

  Identifying copyright restrictions and undertaking due diligence to ensure that they are complied with may be more difficult with born-digital material than in the analog world especially considering the ease with which material can be distributed or collected, with the true origins of the file becoming lost as files are renamed or otherwise lose their context and provenance. Institutions need to be clear about the reason or justification for any restrictions on access to or copying of content for copyright reasons and to provide a mechanism for individuals or organizations to request that material accessible online is taken down due to a breach of copyright.

- Add and edit descriptive metadata.

  Metadata created and/or edited by the archivist will supplement technical metadata captured with the files, to describe their content. Descriptions should be applied at the appropriate level to provide the user with sufficient information within the bounds of what is feasible and scaleable. Descriptive metadata will also represent the intellectual arrangement as discussed at the beginning of this section.

- Record actions and criteria or methods applied.

  As a default, the processing plan will also serve as a record of the work undertaken and the criteria or methods used. However, where there are differences between the plan and implementation, which are likely within an iterative process, it is important to record work as actually carried out rather than merely as planned. This is particularly the case with reference to criteria for appraisal and access restrictions and the rationale for intellectual arrangement.

OBJECTIVE 4: Post-processing steps

Outcome: *The main outcomes of this process tend to be administrative or procedural.*

Decision points

There are no real decision points in this process, rather completion of tasks and processes that signal the end of arrangement and description.

Tasks

- Remove processing restriction (if necessary).

  If the institution has allowed access to the material as soon as it has been accessioned, it may have been necessary to place a restriction on access while the collection was being processed. Upon

completion of processing by the institution, this restriction can be lifted and details relating to how the material can be accessed should be updated.

- Deliver content and metadata to storage (preservation).
- Deliver content and metadata to delivery environment.

The final step in the arrangement and description element of the framework is the effective hand-over, to the storage or delivery environment, according to the institution's infrastructure so that discovery and access can be implemented.

4. Discovery and Access

Definition and Scope

Discovery and Access: the systems and workflows that make material, and the metadata that support it, available to users while ensuring compliance with any access restrictions. The process of discovery and access requires some action on the part of individual users — for example carrying out a search or requesting an item.

Discovery and access in the AIMS framework encompasses the "access" functional component of an OAIS archive: the processes and services by which users locate, request, and receive delivery of items residing in the archival store.[44]

A key difference with born-digital as opposed to paper-based material is that access does not imply a user consulting the original unique object itself. Instead, one or more digital copies are generated at some point in the workflow, for users to access.

Preface

Discovery and access workflows are the final step in the stewardship of born-digital materials in the AIMS Framework. These workflows are shaped by the needs of user communities, but also need to be carried out with regard to legal and ethical issues relating to the material and the information contained within it.

Placing born-digital archives online and making them freely available enables institutions to make both metadata and content easy to discover. This free access also significantly increases the risk of misuse or abuse of copyrighted or sensitive information. In the traditional paradigm, wherein a researcher visits the reading room to request and view materials, institutions rely on personal processes (remote or face to face) to ensure users understand the implication of using materials. However, with digital formats, there may be no logistical need for the institution to be involved in any direct way in users' interactions with the material. This is an unprecedented scenario for institutions, where, depending on the situation, the user potentially has no engagement with an archivist. This creates an environment of isolation for both researcher and archives staff. The institution then suffers from a decline in familiarity with its user base, ability to provide key services, and ability to carry out its duty to the donor and owners of copyright and other intellectual property rights. The researcher on the other hand is increasingly responsible for ensuring that they, on their own, can find relevant records and engage with them in a legal and ethical manner.

[44] Consultative Committee for Space Data Systems. (2002). *Reference Model for an Open Archival Information System (OAIS)*. CCSDS 650.0-B-1. Washington, D.C.: CCSDS. Retrieved from http://public.ccsds.org/publications/archive/650x0b1.PDF

The difficulty of assessing access and use restrictions for records increases with the massive scale of born-digital collections. When combined with the growing professional trend of MPLP practices,[45] accurately and completely ensuring that restricted materials are not accessible becomes more arduous. The fact that making previously unpublished material freely and universally available on-line is regarded in the UK as constituting publication[46] creates a further obstacle — perhaps the most significant one — to fulfilling the potential for access that born-digital material presents. In short, completely unrestricted access to born-digital collections may not be possible, or legally or ethically preferred in many cases.

Keys to Success

When planning for discovery and access of born-digital material, there are several key factors crucial to success. The first is trust (by both donors and users) that the institution, its systems, and processes ensure compliance with intellectual property and confidentiality requirements and also maintain good data management practices. The establishment of this trust stretches back to the first steps in the AIMS Framework: establishing the relationship with the donor or creator during collection development. Trust ensures that the institution can continue to cultivate relationships with donors and users that are vital to the fulfillment of its mission. The institution can gain this trust through clear statements regarding usage rights, clear and effective policies on restriction and data curation, and demonstrating a working system of access restriction and long-term preservation. The institution can also undertake to gain third-party accreditation such as the Trusted Digital Repository certification or the Data Seal of Approval.[47]

For the institution's part, success can be measured in terms of practices and policies that render it possible to build this trust with users and donors and to provide reasonable access to materials. For many, this means a policy of risk management in its approach to providing services. Since paper-based practices do not scale with the increasing volume of born-digital materials collected, the institution needs to spend more time basing decisions on the level of risk associated with an activity, rather than waiting until they are sure of the outcome. This kind of risk-based assessment is evident in decisions to make material available online when, for example,

[45] Greene, M. A., & Meissner, D. (2005). More product, less process: Revamping traditional archival processing. *American Archivist,* 68(2), 208-263.

[46] Padfield, Tim. *Copyright for Archivists and Records Managers*, Third Edition, London 2007, pp93-96.

[47] The criteria for a Trusted Digital Repository certification are outlined in the Trustworthy Repositories Audit & Certification: Criteria and Checklist co-authored by the Research Libraries Group and the National Archives and Records Administration. The checklist can be found at: http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf. The Data Seal of Approval was developed by Data Archiving and Networked Services (DANS), a Dutch organization for research data in the social sciences and humanities. The Seal of Approval is granted by an international Data Seal of Approval Board based on sixteen guidelines found at: http://www.datasealofapproval.org/sites/default/files/DSA%20booklet_2-0_engels_mei2010.pdf.

copyright ownership is uncertain. The benefit of increased access in this case must be weighed against the risk of copyright infringement.[48]

In the above, and in other cases, decision-making will be an iterative process. Since materials in hybrid or solely born-digital collections may not be as fully processed and appraised as they were in the past, understanding content and user requirements as well as the access constraints and restrictions are likely to increase over time. This means that access to materials may change as more information that affects usage policies comes to light. Equally, new techniques for searching the content of digital objects may make the identification of sensitive or protected material easier and more efficient.

Discovery and access are not possible without completion of the preceding steps described in this model. The outcomes of those steps have a significant impact on what is either appropriate or achievable in terms of discovery and access. It is therefore crucial to consider issues relating to discovery and access as early as possible — beginning with the collection development phase — and continuing to update and revise plans as work on the collection progresses.

Specific activities undertaken in earlier parts of the model that may affect discovery and access include:

- **Collection Development**

  Decisions on appropriate access processes, and the extent to which access to some or all of the material can be granted, require agreement between the donor and institution during this initial phase. The donor may also be able to provide guidance on record series or formats which may contain sensitive information.

- **Accessioning**

  Information gathered during accessioning about the nature of the material is vital to determine access policies and methods for different material types. Relevant information from agreements made with the donor about access and restrictions should be captured and retained within collection-level metadata, and as much technical metadata as possible should be extracted or created from the files. For example, metadata such as file format and size may be vital for decisions about access.

- **Arrangement and Description**

  The level at which arrangement and description (or other, more minimal processing) is undertaken (such as at fonds or series) may determine the extent of access. The scope and application of access restrictions may be dependent on the granularity of metadata available or created: the more granular the metadata, the more precise (and less universal) the access restrictions may need to be. Any

[48] See The Society of American Archivists recent endorsement of the recommendations of OCLC Research in their document entitled "Well-intentioned practice for putting digitized collections of unpublished materials online." presents a formalization of this risk-management approach. Both SAA's statement and OCLC's Document can be found at http://www2.archivists.org/groups/intellectual-property-working-group/well-intentioned-practice-for-putting-digitized-collections-of-unpublished-materials-. . See also: McLeod, R. (2008, September). Risk assessment : Using a risk based approach to prioritise handheld digital information. In Fifth International Conference on Preservation of Digital Objects, London, UK. Retrieved from http://www.bl.uk/ipres2008/presentations_day1/20_McLeod.pdf

metadata created before access is granted should be checked and verified for accuracy and fit with institutional policies before proceeding to ensure that the process can continue smoothly. This might include consultation with the creator and/or donor of the material.

Objectives

OBJECTIVE 1: Select and implement access model(s) based on requirements of the collection and of the designated user community/ies

***Outcome**: After analyzing significant properties of content, user requirements, and constraints related to material format and content, an appropriate access model is chosen for the material*

There are three basic requirements which must be met when deciding how to provide access. The first and most basic step is to make material available to user communities by creating a system wherein material can be stored and retrieved. The second requirement is to apply appropriate access restrictions and processes to protect confidential material, copyright, and other intellectual property rights. Thirdly, it is desirable, as far as possible, to provide access to material in a format and/or environment that presents the significant properties that user communities require for their research. All of these factors lead to the determination of an "access model" or the suite of provisions and services that will allow users access to content.

There is a clear link between access models (for access to content) and discovery models (for access to metadata) Although from the users' point of view, discovery usually comes before access; from the archivist's point of view, decisions about access models need to be in place before material is made discoverable. Models for access may be limited by the functionality of an institution's discovery or storage environment or format and content (such as third party privacy issues inherent in email (see "Visualizing Email Access: MUSE" on pg. 46). Access requirements and restrictions may equally influence – or even dictate – decisions about discovery.

The access model must define several characteristics of access including:

- Data format: original, migrated, emulated, disk image
- User location: onsite access (in-person) vs. remote access
- Permission: authenticated access vs. non-authenticated (open) access
- Transfer: physical or off-line access vs. online access
- Creation of access derivative: dynamic in response to request vs. stored derivative access copy
- Restriction level: discover, view, render, or download

A further exploration of these models is included in *Chapter 4, Table 1: Access Models*.

In general, manual and on-site processes retain most control, are risk-averse and require the least technical infrastructure. They also may be the only option when specialized or obsolete software is required to present significant properties. However in many cases they offer the least potential for user access. Online transfer of material — particularly open access with no restriction — is likely to provide the richest user experience but carries risks that require robust technical infrastructure and significant planning.

Decision Points

The primary activity of this objective is decision-making. This requires that information is gathered from previous activities as well as new information-gathering activities to be undertaken. The determination of the appropriate access model for the material in question involves:

- Evaluation of the significant properties of content.
- Analysis of user requirements for discovery and access.
- Exploration of the constraints of the material format (for example, software requirements) or content (the redaction or restriction of sensitive or private material).
- Understanding of institutional infrastructure and support including currently available content management and preservation storage systems.

The decision-making process is iterative in many ways. Firstly, an "ideal" access model may be envisaged during collection development, but amended for practical reasons once the outcomes of arrangement, description, and appraisal are known. Secondly, the access model may be specific to a particular usage instance and may vary for materials within a particular collection or for different groups of users. Thirdly, although one access model may be chosen as a default before a request for access is made, situational decisions may be made after an individual request is received.

Visualizing Email Access: MUSE

Sudheendra Hangal
PhD candidate, Computer Science Department
Stanford University

Email archives are an excellent resource for researchers because they silently record many of the donor's actions and thoughts, forming a passively acquired life-log of his or her everyday activity. However, given the vast amount of data in a long-term email archive, (often running into tens or hundreds of thousands of messages), researchers need good tools to explore and interact with the contents of the archive.

MUSE (Memories USing Email) is a research system designed at Stanford specifically for this task. MUSE uses data mining and text analysis techniques to analyze the email archive and generate cues to messages likely to be of interest; these cues serve as entry points into a browsing interface that supports faceted navigation and rapid skimming of messages. MUSE takes care of data cleaning tasks such as removing duplicate messages, and resolving situations like the same person having multiple email addresses. It can automatically identify important groups of people in the email archives based on co-recipiency patterns, so a researcher could explore all messages involving a particular group of people. It also provides a chronological summary overview of the archive, by identifying the statistically most significant terms in the archive on a monthly basis.

Additionally, MUSE incorporates sentiment analysis techniques to identify messages likely to be of interest – e.g. one use of these techniques would be to adjust the terms to scan for sensitive material in the archive. It can also provide a quick way to scan all the image attachments. While MUSE was originally intended for end users to browse their own long-term archives, the developers of MUSE are working with Stanford libraries to add features useful for archivists or researchers. In addition, it may be useful for donors to use MUSE themselves to clean up their archives before donation. More information about MUSE can be found at: http://mobisocial.stanford.edu/papers/uist11m.pdf; the current prototype of MUSE can be downloaded and used from http://mobisocial.stanford.edu/muse.

Tasks

- Gather relevant data regarding user community.

  This may be performed based on institutional policy or by undertaking research into the user community or even multiple communities. This research should be to uncover the requirements for these users in order to answer questions like: What level of metadata would be appropriate? Would technical metadata such as bit-depth for images, or compression rate for audio files be necessary? Will users be able to, or need to, use original file formats? The answers to many of these questions may vary from user to user, but general assumptions can usually be made regarding some details.

- Analyze file technical metadata and determine handling based on institutional policies and capacities.

  Generally speaking, institutions will want to make all material as accessible as possible, no matter what the file format or other specific details. However, this may not be technically feasible in all cases. Policies should be in place about the treatment of specific formats. For example, an institution may decide to migrate all text documents to PDF for access rather than managing an environment to provide access to multiple word processing software packages in the reading room.

- Take account of confidentiality-, donor-, and copyright-based restrictions indicated in metadata Specifically, address the following:
  - What restrictions have been placed by donor or related parties?
  - Which materials contain confidential information? How comprehensive is your knowledge of this issue?
  - Which materials are subject to copyright and data protection restrictions?
  - What access might still be possible given the above restraints? How can you adapt your access models to facilitate this?

- Assess institutional infrastructure for access to materials (i.e. catalogue, online finding aid, repository, etc.) to determine workflow for provision of access:
  - What is your discovery system? Are you able to create, edit, or synchronize metadata appropriately?
  - Is all required metadata present to create a record for the collection in your access and discovery system (i.e. catalogue record, online finding aid, etc.)?
  - What is your storage environment for access derivatives? Are you or are users able to transfer and retrieve files from the storage environment as appropriate?
  - Is there a security layer linked to the storage environment that keeps restricted materials from public view? Is an authentication system which regulates specific levels of use to individuals needed and in place? Is the workflow linking these layers automated?

- Determine the status of the content that you are able to provide access to, and how you enable users to understand what they are seeing:
  - Content in its original format
  - Content migrated to an alternative format, that is more stable for preservation and/or more accessible without specialized or obsolete software

- Content within a storage and access environment that emulates that of the creator of the material
- The original and migrated formats to be provided are determined by institutional policies on software support and also by digital preservation policies
- Decide on an access model based on the gathered information on content, institutional policies, and infrastructure

Access models (Table 1)

Access element	Access options – description	Factors in decision
Data format	• **Original data/original media** • **Emulated environment** • **Migrated version** • **Disk Image**	The decision regarding the format in which to provide content should be based on the significant properties of the content in question. These significant properties can be influenced by a specific user's research need but will, in general, address most typical research needs. Each type of data format brings with it difficulties for implementation. Original data is the most archivally sound and complete, but formats may be obsolete or require software or even hardware that may not be available to the user. The user on their own may not have the capacity to deal with the material in its original format, so the institution needs to consider what type of support they are willing to offer in accessing the data. In addition, while many users may not find anything remarkable about the original media carrier, others may be very interested in labels, decorations, or other modifications to media. This adds a level of physical preservation that the archives may not have the resources to support. Photographic images of the original *may* be an acceptable substitute. An emulated environment requires no transformation of the original data, and therefore no loss of data, but significant work in recreating the original display environment. A completely faithful recreation of the original environment may not be possible. Migration on the other hand, transforms the data itself to work with newer software and operating systems. The potential for loss of secondary characteristics of the data is enormous, for example formatting, "look and feel," interactivity, and others. These may present an unacceptable loss to some researchers.[46] Finally, a version of the original data captured via a disk image could be used. While this would present the complete original data to the user, with all the benefits identified therein, it adds another layer of complexity to access to that data, since not only the original data format has to be supported, but also the disk image format. In addition, if the image was a forensic copy of an entire disk, the institution is opening itself to the risk of exposing potentially private, sensitive, or copyrighted information.

[46] The Planets Framework provides a methodology identify requirements and evaluate solutions to ensure reliability, integrity and usability of migrated data. http://www.openplanetsfoundation.org/
Plato is a planning tool developed by the Planets programme that implements the planning process and is integrated into a Digital Repository to create access copies of born-digital files through the use of third-party migration or emulation tools and utilities. http://www.ifs.tuwien.ac.at/dp/plato/

Access element	Access options – description	Factors in decision
User location	**In-person/on-site** This enables the digital material to remain within a controlled environment, which the user must visit to gain access. If a high level of control is needed, access should be provided via a standalone machine, or one that is connected to a limited network (e.g., intranet). This may also require blocks on unauthorized copying by the user (e.g., disable or write-protect USB ports) **Remote access** This takes digital material (i.e., a delivery copy) out of the controlled environment of the archival institution. The level of control maintained depends on how closely the user is defined and whether authenticated or not, how closely the material to be accessed is defined, and the method for the transfer of material. In cases where copyright of the material needs to be protected, a risk management approach should be used. For material that contains confidential information, remote access will present too high a risk in all but exceptional circumstances and in the UK may contravene the Data Protection Act. It is important to note that remote access does not necessarily mean online access. A user may remotely request that a copy of a disk be created and sent to them. On the other hand, data may be available online, but a user might be required to manually connect to a campus LAN or through a physical IP address to obtain access. While it is true that these are unusual cases, it is helpful to distinguish the two for finer definition of the access model.	In most cases the user community is likely to prefer remote access. However there will be cases (collections, record series or users) where on-site access is necessary or desirable, such as: • the nature of the records requires strict guarantees on compliance with access restrictions, particularly relating to copyright or confidentiality – this may be difficult to achieve if content is accessed remotely • there is a requirement for the archive institution to authenticate users' identity, and mechanisms are not available for doing this remotely • users need to access material within an emulation environment, and/or using specialist software not widely available • users need to consult paper-based and digital material together, within a hybrid collection • firewalls or user's connectivity prevents transfer of large files or large quantities of files

Access element	Access options – description	Factors in decision
User authentication	**Non-authenticated** Access to born-digital material should not present a risk to its long-term preservation in the way that it does with traditional archives, as the user should be given access to presentation copy of items rather than unique originals. Therefore, in the case of material with no access or copyright restrictions open access requiring no user authentication may be employed, although some information may be gathered for monitoring or advocacy purposes if deemed appropriate. **Authenticated** On-site authentication can be done using existing systems and processes for users visiting to consult paper-based material. The technical requirements for remote authentication depend on the transfer method and level of control required by the content of the material. Different stages of the authentication process may be manual or automated, depending on the status of the user and the functionality available within institutional systems. They may include: • identification and authentication of IP or e-mail address • security layer (requiring user login) • authentication layer (recognizing user login) • ability to register new users with appropriate authentication and access rights (e.g., content-specific, time limited) • Machine-actionable metadata relating to access status	User authentication may be required to fulfill key principles of the stewardship of born-digital archives, in particular: • to record, audit or monitor access and use • to ensure material is accessed only by individuals with necessary authority or appropriate credentials • to receive guarantees from user relating to compliance with copyright or other restrictions on use If no user authentication is needed, it may still be desirable to require user registration in order to gather information about user communities to generate evidence or feedback about the current process for consideration by the institution.

Access element	Access options – description	Factors in decision
Transfer method	The options used for methods to transfer material to the user will depend on the nature of the storage environment and retrieval systems within the archival institution, policies on use of external tools and services, and the degree of control required. With any off-line or physical transfer method it is important to ensure that data (content and metadata) is packaged to ensure that nothing is lost or altered in the transfer process. **Physical, off-line transfer** The simplest retrieval and transfer methods have fewer technical requirements, but may be more labor intensive, depending on the quantity of files and granularity of retrieval or selection. On-site users can consult material in the search room via a non-networked machine with any specialist software required, or copied to disk for use on their own laptop. This offers scope for preventing or monitoring copying of material. Off-site users can be sent discs through postal or courier services **Online** Online access methods can be on- or off-site and can be provisioned to a specific set of files for a one-time use or a continuously granted to a more general range of material. Options include: • On-site access via dedicated machine linked to institution's local network. • Transfer of specific files via e-mail, intranet, or web-based file sharing tool (e.g., Dropbox or FTP). A risk management approach should be used for use of third-party web-based file sharing tools and services. • Remote access to specific set of individual files or more general range of material via digital repository.	Methods used for retrieval and transfer will depend on a balance of three key factors: • Degree of control required: physical and manual methods give archivists more control over access to material and (in the case of on-site visitors) enable them to prevent or monitor copying of material. • Resources available: manual processes (whether physical or on-line) are likely require more staff time in dealing with individual requests, but require less investment in technical infrastructure to provide the same level of control over access to material • Technical infrastructure: automated on-line access requires an institutional digital repository with systems to identify closed and open content and manage access appropriately

Access element	Access options – description	Factors in decision
Retrieval / generation of content	The process by which the content is generated for the user may fulfill a "just in case" role (Static) or a "just in time" one (Dynamic) **Static (S)** Access versions of content are generated without a user request being made. Access versions are therefore the same for all users, both in terms of the nature of the content and the application of access restrictions. Static access models include: • PC or discs with static content issued to users in search room • Material in institutional repository open to general public access via online link **Dynamic (D)** Access versions of content are generated in response to a specific user request. If required, or applicable, this gives the potential for users to access different versions with specific access permissions. Dynamic access models include: • Material retrieved and copied for specific user • DIP generated dynamically by repository in response to an access request, potentially enabling format(s) for material to be specified	The decision to dynamically generate content may seem more labor intensive at first glance, however the large scale creation of derivatives, especially in normalized or migrated formats, is not trivial. This is especially true of content that needs to be surveyed for sensitive or copyrighted material before it can be made available. As an alternative to doing this work ahead of time, an institution may decide to generate access copies as requests are made. As with the paper environment, MPLP advocates strategies that delay detailed processing (to identify restricted material, for example) until a user request has been made. This gives the institution the added flexibility of creating access derivatives that meet users specifications for data format (for example, providing JPEG2000 derivatives of TIF masters instead of the typical JPEG file). However, with very large quantities of material or requests, dynamic creation of derivative files by staff members is not feasible. Increasingly repository systems can generate access derivatives dynamically, but the number of data formats for which this is possible is never likely to be large. Most institutions will adopt a hybrid of these approaches and have large quantities of access derivatives of low-risk materials in common data formats while continuing to dynamically generate access copies for special requests and more sensitive materials.
Restriction level	The level of restriction applied is determined during processing. The levels include: **Discover** Items may be identified by a search for metadata. **View** Metadata may be viewed. **Render** Browser-renderable representations of an asset can be displayed. **Download** Associated files can be downloaded	These options were derived from the functional requirements developed for a tool to facilitate Arrangement and Description (see *Appendix H.1*). Although they do blend access and discovery, which is described more fully in the next objective, they are included in the access model since they have an effect on the access granted to content.

Publication Pathway and Discovery and Access at the Bodleian Library

Susan Thomas, Bodleian Library

The development of the Bodleian's publication pathway for digital archives has been driven by a number of factors, each imposing its own pressures.

At a high-level, we wish to avoid a boutique-like approach and move straight to a workable framework that will provide a baseline minimum for access and discovery across all archives containing digital materials. This minimum-level applies irrespective of the significance of the archive or it's creator, and without regard to the quantity of the digital material. As this is written, collections ready for some form of dissemination to users include archives with a handful of digital items, and archives with thousands.

Our baseline minimum

We have adopted a browser-based discovery environment (currently Drupal-based) and a migration approach to providing access in that environment. Exploration of emulation-based access is deferred until we are satisfied that our minimum access and discovery systems are sufficiently usable, and that we have a willing depositor with a digital archive that would benefit from the application of emulation techniques.

Versions of our collections

Based on metadata supplied by the processing archivist, our publication pathway may create two dissemination versions of a collection ready for publication. The first version is intended for reading room use only, deployed via a private network to dedicated locked-down clients; the second version is destined for online access. The processing archivist will have assigned dates for release of material into each of these environments, weighing ethical and legal issues including data protection and intellectual property rights. If material has not been cleared for release into a particular environment, then that material, and metadata about it, will not be available in that environment.

Prioritizing development

The development of the migration paths, and associated services, which are used during our collection building process have been influenced by the processed collections which we are pushing through the publication process. To date, much of these have consisted of legacy word-processing formats, though still image, audio and moving image items are now entering the publication pathway in larger numbers. Likewise, the decision to create static — rather than dynamic — versions of items for access is driven by the sometimes complicated migration pathways, and the need for quality assurance work. It is possible that we will look to use migration-on-demand methods for some kinds of material in future.

The prioritization of discovery tools has also been driven by the collections in the publication queue, with much of the initial focus being on the textual content that has pre-dominated, with keyword clouds and full-text search provided for these materials. This is combined with metadata search and browsing, which provides at least some level of access to non-textual materials.

It is still early days for our work in this area, and each collection brings fresh challenges that expand the capabilities of our publication pathway. We still have much to do if we are to meet our users' expectations.

OBJECTIVE 2: Make material discoverable by designated user community

***Outcome**: Designated user community can discover material through metadata relating to material at collection and/or series and/or item level and, in some cases, through the content of the material itself. Where*

metadata or content contains restricted information, this information is not available for search, index, or access.

Thinking about discovery as a separate stage in the workflow enables us to consider the process from the user's point of view and therefore to consider how the processes of discovery and access are linked. As this objective is concerned only with the discoverability of material, not the accessibility of the material itself, it is primarily concerned with metadata and in some cases the searchable text of the content itself. Not all digital objects will have searchable content because the data itself does not contain text, or because that content is restricted. Metadata will most likely still be created for collections with searchable content as the result of full arrangement and description or accessioning processes. The scale of material will likely make description at the item level impossible and therefore descriptive metadata will most likely be created at the collection, series/fond, or other aggregate level. Technical metadata created through automated methods such as the use of metadata extraction tools can and should be provided at the item level where appropriate.

Some will advocate adopting an MPLP approach and making content itself available for search together with just the minimal metadata derived from accessioning. This is a risky proposition precisely because of the issues related to private and sensitive data and copyright discussed at the beginning of this section. However, in collections with minimal risk of this type of data, enhancing discovery through full-text searching is a distinct advantage of born-digital material.

Just as there are models of access to material, so too are there models of discovery. These are explained more fully in *Chapter 4, Table 2: Discovery Model*, but the basic types are:

- Discovery through metadata only
- Discovery through content only (full-text search)
- Discovery through both metadata and content

Decision points

As with the previous objective, the discovery objective hinges on the selection of an appropriate discovery model. The decision should be based on the available resources at the institution as well as the nature of the material. For example, an image collection is obviously not appropriate for discovery through full-text search. As another example, an institution with a staff of one or two probably won't have the resources to develop a sophisticated content repository with robust item-level descriptive metadata. A minimal level of access could be facilitated at most institutions by providing descriptive metadata about born-digital material in the same catalogs or finding aids that contain metadata for discovery of paper-based materials. Discovery though content as well as metadata, and the use of technical metadata for discovery, can be of secondary consideration in these cases.

Generally speaking, the more data that is available for search, the better the discovery will be. Most institutions will opt to make as much full-text and metadata available as they can feasibly provide. The work of being a good steward of born-digital materials in this area then is in balancing the needs for discovery against the resources available, even if it means being unable to provide "ideal" access.

Tasks

- Review system(s) interfaces to determine whether all metadata and discovery functions (including search, browse, and filter) needed by user community are supported. Amend as appropriate.
- Review system(s) indexing rules to determine whether all metadata is appropriately indexed. Amend as appropriate.
- If processing and access are undertaken using separate tools, test and establish methods for data transfer or synchronization between the two.
- Publish searchable data (from content or metadata) to appropriate system.
- Provide a means for the user to progress from searching to retrieving the content based on the chosen access model, such as:
  - Through an automated authentication.
  - Through a request for access.
  - Through an onsite visit.
- Create guides to describe how to access born-digital material and to help users understand the nature of the content and the possible implications of migration and normalization processes.

Discovery models (Table 2)

Discovery Type	Description	Factors influencing Decision
Discovery through metadata only	Discovery through metadata only means that the institution provides adequate information about the collection, at the appropriate level of granularity, to lead users to content. This metadata may be searchable through an online system that the institution manages, or could be found on static web pages that are indexed and search by a web search engine. Access to the content then may be through links or instructions to request access in person, or through some other means. Examples: • Collection guide finding aid or catalogue (EAD or webpage) • Metadata in a searchable online system	Discovery through metadata is the traditional model for information discovery and therefore is often the easiest to implement. Simply posting a static html collection guide online enables users to discover the existence of content through any web search engine. Other institutions may have their own searchable database of guides, or may have a digital object repository with non-archival metadata. While this system is the most traditional, it does not have the advantages of full-text search and the creation of high-quality descriptive metadata is resource-intensive.
Discovery through content only	Discovery through content only means that no metadata is exposed for users to discover, instead the full-text of the content is indexed and searchable. Having content available for searching does not necessarily mean that the content is accessible for viewing. The Google Book Search "snippet" approach[47] is an example of a content-based discovery system that does not necessarily provide full access to searchable content.	Access through the content itself seems to many to be the ideal way to provide access: there are low overhead costs for the institution and little need to interpret the information that is already found in the source. However, the process is not without drawbacks and risks. As has been discussed, the exposure of content increases risk of exposing sensitive, private, or copyrighted data. In addition, full-text may not be the best text for discovery. The synthesis of concepts and controlled vocabulary found in metadata can often make discovery of complex objects easier. Technical implementation of full-text search may also be more difficult than metadata-based discovery if a searchable index of the full-text content needs to be created. However, simply placing the text online will allow for web search engines to provide some simple level of access.
Discovery through content plus metadata	In the hybrid approach, both metadata and the full-text of content are searched. Since content is available, the same stipulations regarding restricted content are necessary as with the discovery through content only system.	This approach offers the advantages of both systems. In fact it may slightly mitigate some of the work of metadata creation if minimal metadata only is used to enhance full-text search. The technical bar for implementing this type of access may be higher than the others due to a lack of systems on the market that enable this feature in the archival environment. In addition, the combination in a search environment of the "about-ness" of metadata with the "of-ness" of content requires more sophisticated handling of search results.

[47] http://books.google.com/googlebooks/screenshots.html#snippetview

OBJECTIVE 3: Provision of access to content when dissemination requests are received.

> ***Outcome**: Individual access requests are received and processed. The user understands the archival and technical context and status of the content and restrictions are fully complied with. The user fully understands their responsibilities and limitations based on copyright or data protection laws. Access details are recorded for audit trail where required by institution and/or depositor/donor.*

This objective relates to the provisioning of access to the collection material itself. General policies for access methods must be in place and a discrete action initiated by a user request for material prior to carrying out this objective. All tasks under this objective may be carried out by a repository system without any interaction by staff. These tasks are enumerated here, however, to provide a model of how access is actually provisioned.

Decision points

As this objective is far more action oriented than the previous two, decision points here all relate to specific activities. Many of the decision points and tasks below are taken from Section 4.1.17 of *Reference Model for an Open Archival Information System (OAIS)* (2002).[48] Although these are intended to be written into the handling of traffic in a repository system, it is useful to think about them outside of this closed system. In institutions that do not use a repository system, but instead are combining a series of manual workflows to provide access to materials, these activities will be performed by staff.

The major area of decision takes place when a dissemination request is received. The following activities will be influenced by the generic or ideal access model that has been adopted. The archivist must decide if that model of access will be followed or if the request merits special handling. Once this has been determined the archivist can continue to process the request.

Tasks

- Receive dissemination request and identify appropriate access model:
  - Generic or ideal access model
  - Access type for user category
  - Customized as per request
- Determine if resources are available
  - Identify access restrictions / status of material
- Assure that user is authorized to access where required
  - Obtain declarations from user
  - Obtain user authentication
  - Acknowledge user authentication
  - Give permission for access, within defined parameters
- If access to content is dynamic:

[48] Consultative Committee for Space Data Systems. (2002). *Reference Model for an Open Archival Information System (OAIS)*. CCSDS 650.0-B-1. Washington, D.C.: CCSDS. Retrieved from http://public.ccsds.org/publications/archive/650x0b1.PDF

  - Retrieve Archival Information Package (AIP)
  - Generate Dissemination Information Package (DIP)
  - Determine delivery format and environment based on prior decisions about access models
  - Undertake other special processing as needed
- If access to content is static:
  - Retrieve DIP
- Deliver DIP
- Record access details if required

OBJECTIVE 4: Gather information for future decision-making purposes

> ***Outcome****: Data gathered from audit trail, usage statistics, and other user feedback mechanisms is used to continuously improve discovery and access.*

Once discovery and access procedures and workflows are in place, information should be continuously gathered to regularly improve discovery and access to the materials. As was discussed earlier, the nature of increasing remote, online access to materials isolates the institution from its users. In response, institutions must gather information in other ways about user interactions and services provided to continuously improve them. We no longer have the luxury of anecdotal feedback and evaluation of services we might have received at the search room desk. If we are to continue to understand our users and to provide them with valuable services, we must be proactive in seeking out their input.

Decision Points

There are many ways in which libraries and archives gather information about their users and usage: circulation statistics, user studies, web usability testing, and surveys are just a few. The archivist needs to evaluate these methods to determine how and where they can be enhanced, if necessary. Methods for gathering and analyzing information should be documented, and once that analysis is done that information should be disseminated to staff and possibly even users for feedback. The process of gathering feedback is iterative and evolving.

The initiative to gather feedback is, however, only as worthwhile as the conviction to do something with it. Feedback can indicate potentially large and complex changes to workflow, and sometimes practicality will win out. However, the institution will be strengthened by a willingness to look at itself critically and be open to changes to procedure. Aligning their work with the services their users actually ask for will be rewarded through continued and perhaps increased usage.

Tasks

- Ensure that methods are in place to track usage of materials through means such as
  - Web analytics and transaction logs
  - Registration rolls or other reading room tracking measures
  - Archiving of user requests
- Solicit feedback from users through:
  - Online comment or feedback forms
  - Usability tests
  - Focus groups
- Analyze available information (statistics, user feedback) on a regular basis to determine patterns and needs
- Disseminate findings as appropriate
- Initiate changes in service based on findings

Conclusions

The AIMS Framework might be characterized as much by what is left out as what is covered. Far from a criticism, the partners hope that, by making definitions of what are, and by extension what are not, core archival functions when dealing with born-digital materials, they are helping to move the field towards more universal best practices.

However, a number of issues that are tangential to the Framework may have significant implications for its successful use. In many cases, substantial work has been done in these areas and should be consulted along with this document. Particular areas of concern include:

- Digital preservation
- Legal and ethical aspects of acquiring disk images of digital media
- Archival data management and the systems that support it
- Issues surrounding born-digital materials on "legacy" media that has already been physically acquired by an archives
- The need for active engagement with donors and users about the particular issues surrounding born-digital collections
- Making the transition from a specialized project to a continuous service and the re-alignment of institutional priorities to address both paper and born-digital collections
- Appropriate forums for sharing the skills with other institutions

As stated in the *Introduction*, the Framework is not intended to be a best practice recommendation or to replace or supersede work done elsewhere. Rather, it is an addition to the emerging body of research and practice that is informing the evolving archival practice. As archives make the transition from mostly analog collections to those predominantly born-digital, the stewardship of these materials increases the complexity of decisions and highlights the need to develop more collaborative relationships across disciplines. Because of uncertainties about almost every aspect of the stewardship process, curators are faced with new questions and few answers. However, we stewards must not forget that born-digital materials also offer new opportunities for discovery and access. The AIMS project, conceived as an "inter-institutional" initiative, aspires to serve as a model for collaboration as well as a pathway for managing our collections for the 21st century and beyond.