



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.pdf.





Table of Contents

Foreword	i
The AIMS Framework: The Functions of Stewardship	1
1. Collection Development	4
Donors and Trust: University of Hull	8
Enhanced Curation at the British Library	13
2. Accessioning	17
Evolution of Accessioning at University of Hull	21
Project Xanadu: Loss and Recovery	25
3. Arrangement and Description	31
Technical Development: Functional Requirements for Arrangement and Description	33
Arrangement and Description Case Study: The Papers of Stephen Gallagher	36
4. Discovery and Access	44
Visualizing Email Access: MUSE	48
Access models (Table 1)	51
Publication Pathway and Discovery and Access at the Bodleian Library	56
Discovery models (Table 2)	59
Conclusions	63
Appendix A: Glossary	64
Appendix B: Bibliography	72
Appendix C: Contributor Biographies	76
Appendix D: Institutional Summaries and Collection Descriptions	83
1. The University of Hull, University Archives at Hull History Centre	83
2. Stanford University, Stanford University Libraries & Academic Information Resources	86
3. The University of Virginia, Albert and Shirley Small Special Collections Library	91



4. Yale University	93
Appendix E: Sample Processing Plans	96
1. University of Hull: Stephen Gallagher Processing Plan	96
2. Stanford University: Gould Processing Plan	99
3. University of Virginia: Cheuse Papers Processing Plan	104
4. Yale University: Tobin Collection Processing Plan	106
Appendix F: Policies, Templates, Documents, etc.	108
1. AIMS Donor Survey	108
2. University of Hull Accessioning Workflows	113
3. University of Hull Digital Media Photography Form	115
4. University of Hull Insertion Sheet	116
5. Guidelines for Creating Agreements at Stanford University	117
6. Stanford University Processing Workflow	120
7. Beinecke Library Born Digital Archival Acquisition Collection and Accession Guidelines	124
Appendix G: Technical Evaluation and Use	125
1. AccessData FTK3.3	125
2. AccessData FTK Imager 3.0	127
3. Comparison of 5.25" Floppy Disk Drive Solutions	129
4. Karen's Directory Printer (v.5.3.2)	132
5. Curator's Workbench	134
Appendix H: Technical Development	136
1. Functional Requirements for Arrangement and Description	136
2. Rubymatica	167
3. Hypatia	169
Appendix I: Digital Archivist Community	171
1. Born Digital Archives Blog	171
2. Digital Archivist Community Events	174
3. Day of Digital Archives	179
4. Presentations, Conferences, and Publications	181



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 scalable 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 Project¹ 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.

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.uk/>

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.

Archivemata

<http://archivemata.org/wiki>

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

¹ <http://projecthydra.org>



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 1.1*):
<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 1.2*.
- The creation of the Day of Digital Archives project and blog (see *Appendix 1.3*):
<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.1.1*. 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) methodology² 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.

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

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

- **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.



AIMS: An Inter-Institutional Model for Stewardship

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.



I. 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 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 properties⁴ 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

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:"

⁴ For example, at Stanford, when the Peter Koch computer files were acquired by a logical capture, the fonts associated with his InDesign 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.

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.

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 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.

A complete understanding and description of the infrastructure will include:

- storage environment
- equipment for transfer, capture, and quarantine
- 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. 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 users⁹ 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

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

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.

- Ownership of materials (relating to analog material, where content is unique) or exclusivity (relating to born-digital materials, where content is a copy)

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 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

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.

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. 13) 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 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 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

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.

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/>

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 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.

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.

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.

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 MPLP¹⁶ 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.

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

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 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

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 photos 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 within the workflow, assisting with staff training and facilitating the assessment of other software.

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 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.

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).

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.

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 builder²⁰ 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

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

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).

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 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

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

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),

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 Xanadu project to its office space on California Avenue. Before long, however, the arrangement collapsed. The team disintegrated, with Nelson and Gregory regaining control of Xanadu, 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.

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. 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

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

- 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

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.

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:

- 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

- 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

²⁶ Roe, p. 55-56.

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 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.

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.

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).

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

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).

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.

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 recordkeeping 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 Interface³¹
- 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 objects³²

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.

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.

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 project³⁴ 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. 33-34 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.

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.

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.

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

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

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 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

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.

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.

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 collection⁴⁰ 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.

³⁸ See InterPARES I, 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*.

- 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 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.

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.

⁴² <http://www.pedalspreservation.org/>

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 user⁴³ and be capable of being read by humans and processed by machine.

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.

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.

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

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>

responsible for ensuring that they, on their own, can find relevant records and engage with them in a legal and ethical manner.

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 publication⁴⁶ 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

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.

assessment is evident in decisions to make material available online when, for example, 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,

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

the more precise (and less universal) the access restrictions may need to be. Any 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. 48)). 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-reciprocity 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	<ul style="list-style-type: none"> • Original data/original media • Emulated environment • Migrated version • Disk Image 	<p>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.</p> <p>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 <i>may</i> be an acceptable substitute.</p> <p>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.</p> <p>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</p> <p>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.</p>

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	<p>In-person/on-site</p> <p>This enables the digital material to remain within a controlled environment, which the user must visit to gain access.</p> <p>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)</p> <p>Remote access</p> <p>This takes digital material (i.e., a delivery copy) out of the controlled environment of the archival institution.</p> <p>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.</p> <p>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.</p>	<p>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:</p> <ul style="list-style-type: none"> • 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	<p>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.</p> <p>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.</p> <p>Authenticated On-site authentication can be done using existing systems and processes for users visiting to consult paper-based material.</p> <p>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:</p> <ul style="list-style-type: none"> • 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 	<p>User authentication may be required to fulfill key principles of the stewardship of born-digital archives, in particular:</p> <ul style="list-style-type: none"> • 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 <p>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.</p>

Access element	Access options – description	Factors in decision
Transfer method	<p>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.</p> <p>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.</p> <p>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.</p> <p>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</p> <p>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:</p> <ul style="list-style-type: none"> • 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. 	<p>Methods used for retrieval and transfer will depend on a balance of three key factors:</p> <ul style="list-style-type: none"> • 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	<p>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)</p> <p>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.</p> <p>Static access models include:</p> <ul style="list-style-type: none"> • PC or discs with static content issued to users in search room • Material in institutional repository open to general public access via online link <p>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.</p> <p>Dynamic access models include:</p> <ul style="list-style-type: none"> • 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 	<p>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.</p> <p>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).</p> <p>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.</p> <p>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.</p>
Restriction level	<p>The level of restriction applied is determined during processing. The levels include:</p> <p>Discover Items may be identified by a search for metadata.</p> <p>View Metadata may be viewed.</p> <p>Render Browser-renderable representations of an asset can be displayed.</p> <p>Download Associated files can be downloaded</p>	<p>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.</p>

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 its 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 through 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	<p>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.</p> <p>Examples:</p> <ul style="list-style-type: none"> • Collection guide finding aid or catalogue (EAD or webpage) • Metadata in a searchable online system 	<p>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.</p> <p>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.</p>
Discovery through content only	<p>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” approach⁴⁷ is an example of a content-based discovery system that does not necessarily provide full access to searchable content.</p>	<p>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.</p> <p>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.</p> <p>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.</p>
Discovery through content plus metadata	<p>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.</p>	<p>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.</p> <p>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.</p>

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

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>

- Give permission for access, within defined parameters
- If access to content is dynamic:
 - 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.

Appendix A: Glossary

Access model: Combination of environment and processes that enables users to access content and the archival institution to control and/or monitor access as required and to guarantee compliance with access restrictions.

AIP (Archival Information Package): Within the OAIS Reference Model, the AIP is the managed, archival package of digital objects (including both content and metadata) that is stored and preserved by the OAIS system. The AIP is composed of the SIP plus any additional metadata related to any archival processes that were undertaken. See also: SIP, DIP.⁴⁹

API (Application Programming Interface): A collection of computer subroutines published in such a manner that other software can easily invoke the subroutines. API is often synonymous with software library, subroutine library, or module. An API will contain subroutines, functions, methods, and/or classes.⁵⁰

Appraisal: Process of deciding whether or not materials have enduring research value and should therefore be retained. The term selection is also sometimes used for this process.⁵¹

Archivematica: A comprehensive open source digital preservation software system that complies with the ISO-OAIS functional model.⁵²

Archivists' Toolkit (AT): An "open source archival data management system to provide broad, integrated support for the management of archives." It is the result of a collaboration of the University of California San Diego Libraries, the New York University Libraries and the Five Colleges, Inc. Libraries.⁵³

Arrangement: The process of organizing materials with respect to their provenance and original order, to protect their context and to achieve physical and intellectual control over the materials.⁵⁴

⁴⁹ Source: CCSDS Recommendation for an OAIS Reference Model, pg 1-7.

⁵⁰ Source: http://en.wikipedia.org/wiki/Application_programming_interface

⁵¹ Source: http://www.archivists.org/glossary/term_details.asp?DefinitionKey=3

⁵² Source: http://archivematica.org/wiki/index.php?title=Main_Page

⁵³ Source: <http://www.archiviststoolkit.org>

⁵⁴ Source: http://www.archivists.org/glossary/term_details.asp?DefinitionKey=294

Artifactual file: The original file, a copy of which is then processed in working/preservation storage - likely to have its own preservation policy

Audit trail: A means of tracking all the interactions with records within an electronic system so that any access to the system can be documented as it occurs for the purpose of preventing unauthorized actions in relation to the records as well as determining if relevant policies and procedures were followed or, if not, why they were not followed.

Authenticity: The quality of being genuine and free from tampering as well as being what it professes in origin or authorship.⁵⁵

Axiell CALM (Computerisation for Archives, Libraries and Museums): Archives collection management software developed by Axiell. Widely used in the UK.⁵⁶

Catalog: see Finding Aid.

Checksum: A unique numerical signature with a fixed, small length, derived from a file. Used to verify that two copies of a file are identical. Also referred to as a hash value.⁵⁷

Content: The intellectual substance of a document, including text, data, symbols, numerals, images, and sound.⁵⁸

Context: The organizational, functional, and operational circumstances surrounding materials' creation, receipt, storage, or use, and its relationship to other materials.⁵⁹

DACS: Describing Archives: A Content Standard. An output-neutral set of rules for describing archives, personal papers, and manuscript collections, and can be applied to all material types. It is the U.S. implementation of international standards (i.e., ISAD(G) and ISAAR(CPF)) for the description of archival materials and their creators.⁶⁰

Description: The creation of an accurate representation of a unit of archival material by the process of capturing, collating, analyzing, and organizing information that serves to identify archival material and explain the context and records system(s) that produced it.⁶¹

⁵⁵ Source: http://www.archivists.org/glossary/term_details.asp?DefinitionKey=9

⁵⁶ Source: <http://www.axiell.co.uk/calm>

⁵⁷ Source: <http://en.wikipedia.org/wiki/Checksum>

⁵⁸ Source: http://www.archivists.org/glossary/term_details.asp?DefinitionKey=627

⁵⁹ Source: http://www.archivists.org/glossary/term_details.asp?DefinitionKey=103

⁶⁰ Source: <http://www.archivists.org/governance/standards/dacs.asp>

⁶¹ Source: InterPARES 2 Project, Terminology Cross-domain Task Force, pg 5.

Discovery model: Combination of environment, tools, and services which enables users to identify and locate resources of interest to them. At its most basic level, it requires publication or other dissemination of information, and increasingly it also implies on-line availability together with interactive search facilities.

Disk image: A single file or storage device containing the complete contents and structure representing a data storage medium or device, such as a hard drive, tape drive, floppy disk, CD/DVD/BD, or USB flash drive, although an image of an optical disc may be referred to as an optical disk image.⁶²

Dissemination request: A request made of a repository or archive by a user for digital objects or metadata about them.⁶³

DIP (Dissemination Information Package): The package of digital object(s) and metadata that is produced or retrieved by an OAIS system as a result of a dissemination request. See also: SIP, AIP.⁶⁴

Donor: This term is used to denote any person or organization transferring material to an archival institution. The material may be a donation, purchase or deposit (indefinite loan). The term donor is used for convenience to imply any of these scenarios.

DRAMBORA (Digital Repository Audit Methodology Based on Risk Assessment): A methodology for self-assessment, encouraging organizations to establish a comprehensive self-awareness of their objectives, activities and assets before identifying, assessing and managing the risks implicit within their organization. Developed jointly by the Digital Curation Centre (DCC) and Digital Preservation Europe (DPE).⁶⁵

DROID (Digital Record Object Identification): A software tool developed and distributed by the National Archives of the UK to that uses the PRONOM registry to automatically identify file formats.⁶⁶

EAD (Encoded Archival Description): A non-proprietary de facto standard for the encoding of finding aids for use in a networked (online) environment.⁶⁷

Emulation: The reproduction of the behavior and results of obsolete software or systems through the development of new hardware and/or software to allow execution of the old software or systems on future computers.⁶⁸

⁶² Source: http://en.wikipedia.org/wiki/Disk_image

⁶³ Source: CCSDS Recommendation for an OAIS Reference Model, pg 4-11.

⁶⁴ Source: CCSDS Recommendation for an OAIS Reference Model, pg 10.

⁶⁵ Source: <http://www.repositoryaudit.eu/about/>

⁶⁶ Source: <http://droid.sourceforge.net/>

⁶⁷ Source: <http://www.archivists.org/saagroups/ead/aboutEAD.html>

⁶⁸ Source: InterPARES 2 Project Book: Glossary, pg 20.

Fedora (Flexible Extensible Digital Object Repository Architecture): Originally developed by researchers at Cornell University as an architecture for storing, managing, and accessing digital content in the form of digital objects inspired by the Kahn and Wilensky Framework. Fedora implements the Fedora abstractions in a robust open source software system.⁶⁹

Finding aid: A description of records that gives the repository physical and intellectual control over the materials and that assists users to gain access to and understand the materials.⁷⁰

FITS (File Identification Tool Set): Identifies, validates, and extracts technical metadata for various file formats and combines their results into a single XML file. It wraps several third-party open source tools, normalizes and consolidates their output, and reports any errors. FITS was created by the Harvard University Library Office for Information Systems for use in its Digital Repository Service (DRS).⁷¹

File viewer: Application software that presents the data stored in a computer file in a human-friendly form. The file contents are generally displayed on the screen, printed, or read aloud using speech synthesis.⁷²

Forensic disk image / forensic copy: A complete sector-by-sector copy of the source medium and thereby perfectly replicating the structure and contents of a storage device.⁷³

Hybrid collection: A collection consisting of both born-digital and paper-based materials.

Hydra: A multi-institutional collaboration to build a common, open source framework for multi-function, multi-purpose, repository-powered applications. As symbolized by its name, Hydra's architecture reflects a "one body, many heads" design: a robust digital repository (the body) anchors feature-rich applications (the heads), tailored to content-, domain- and institution specific needs and workflows. Hydra's technical framework features the Fedora Repository on the back end, with a front end comprising Ruby on Rails, Blacklight, Solr, and a suite of web services.⁷⁴

Hypatia: A Hydra application (Fedora, Hydra, Solr, Blacklight) that supports the accessioning, arrangement / description, delivery and long term preservation of born digital collections. By using a common set of software tools and APIs, Hypatia will also have features related to access, delivery, authorization, and preservation.⁷⁵

⁶⁹ Source: <http://www.fedora-commons.org/about>

⁷⁰ Source: http://www.archivists.org/glossary/term_details.asp?DefinitionKey=66

⁷¹ Source: <http://code.google.com/p/fits/>

⁷² Source: http://en.wikipedia.org/wiki/File_viewer

⁷³ Source: http://en.wikipedia.org/wiki/Disk_image

⁷⁴ Source: <https://wiki.duraspace.org/display/hydra/The+Hydra+Project>

⁷⁵ Source: <https://wiki.duraspace.org/display/HYPAT/Home>

Ingest: The act of moving a submission information package (SIP) into a digital repository as an archival information package (AIP). Ingest also refers to a specific OAIS entity that contains the services and functions to perform this activity.⁷⁶

Institution: The term used within this paper to describe a collecting repository, record office, or other institution undertaking stewardship of archives.

ISAD(G) (General International Standard Archival Description): A standard to provide general guidance for the preparation of archival descriptions. It is to be used in conjunction with existing national standards or as the basis for the development of national standards. Created by the International Council on Archives (ICA) Sub-Committee on Descriptive Standards (CBPS) (second edition, published 2010).⁷⁷

Logical copy / Logical image: A copy of specific files made from a storage device, retaining their hierarchical organization within directories or folders. The full path of each file is recorded. Deleted files and un-partitioned space are not copied.

METS (Metadata Encoding and Transmission Standard): A standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed as an XML schema.⁷⁸

Migration: The process of converting records to newer formats in order to maintain their compatibility with a newer generation of hardware and/or software computer technology, while leaving intact their intellectual form.

Normalization: The process of creating and/or storing digital documents or other digital objects in a limited number of standardized data or file formats.⁷⁹

OAIS (Open Archival Information System): An archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community. (ISO 14721:2003)⁸⁰

Open source: A development strategy wherein the source materials of an end product are made available. The term is most commonly used in collaborative software development when source code of an application is made available with an application for others to change or improve upon.

Original order: The organization and sequence of records established by the creator of the records.⁸¹

⁷⁶ Source: CCSDS Recommendation for an OAIS Reference Model, pg 11.

⁷⁷ Source: <http://www.ica.org/7102/public-resources/isadg-general-international-standard-archival-description-second-edition.html>

⁷⁸ Source: <http://www.loc.gov/standards/mets/>

⁷⁹ Source: InterPARES 2 Project Book: Glossary, pg 20.

⁸⁰ Source: CCSDS Recommendation for an OAIS Reference Model, pg 1-11.

⁸¹ Source: http://www.archivists.org/glossary/term_details.asp?DefinitionKey=69

PAIMAS (Producer-Archive Interface Methodology Abstract Standard): Standard that covers the first stages of the ingest process defined by OAIS. Identifies and provides a structure for the interactions which take place between an information producer and a deposit archive. (ISO 20652:2006)⁸²

Physical control: The function of tracking the storage of records to ensure that they can be located.⁸³

PREMIS (Preservation Metadata: Implementation Strategies): Core preservation metadata model for organizing and thinking about preservation metadata, defined in the PREMIS Data Dictionary. Documentation includes guidance for local implementations.⁸⁴

PRONOM: a web-based technical registry to support digital preservation services; developed by The National Archives of the United Kingdom. A proposal current underway seeks to bring PRONOM together with the Global Digital Format Registry Project to create a Unified Digital Formats Registry (UDFR). The DROID tool was created to use the PRONOM registry for format identification.⁸⁵

Provenance: The relationship between records and the organizations or individuals that created, accumulated, and/or maintained and used them in the conduct of personal or corporate activity. See also: respect des fonds.⁸⁶

Processing environment: The workspace where accessioning and arrangement and description is undertaken.

Quarantine space: A location where items can be held and isolated in order to mitigate the effect of any contaminants and prevent them from spreading to other materials. In a workflow with born-digital materials, this may involve keeping files on storage media not connected to a server until malware or viruses can be detected and removed.

RAD (Rules for Archival Description): Published by the Canadian Committee on Archival Description. Revised version released in 2008.⁸⁷

Repository: Term used in this paper to refer to the digital repository, not as an alternative term to institution.

⁸² Source: http://www.dcc.ac.uk/resources/standards/diffuse/show?standard_id=154

⁸³ Source: http://www.archivists.org/glossary/term_details.asp?DefinitionKey=978

⁸⁴ Source: <http://www.loc.gov/standards/premis/tutorials.html>

⁸⁵ Source: <http://www.nationalarchives.gov.uk/PRONOM/>

⁸⁶ Source: <http://archives.un.org/unarms/en/unrecordsmgmt/unrecordsresources/glossaryofrecordkp.html>

⁸⁷ Source: <http://www.cdncouncilarchives.ca/archdesrules.html>

Respect des fonds: The principle that the records created, accumulated, assembled, and/or maintained and used by an organization or individual must be kept together in their original order if it exists or has been maintained and not be mixed or combined with the records of another individual or corporate body.⁸⁸

Ruby on Rails: An general, object-oriented open-source programming language. Ruby is tightly integrated with a web application framework called Rails. Ruby on Rails is part of the stack of technologies used in the Hydra and Hypatia projects.⁸⁹

Series: A group of records based on a file system or maintained as a unit because the records result from the same function or activity, have a particular form, or have some other relationship resulting from their creation, accumulation, or use.⁹⁰

Significant properties: Significant properties, also referred to as “significant characteristics” or “essence”, are essential attributes of a digital object which affect its appearance, behavior, quality and usability. They can be grouped into categories such as content, context (metadata), appearance (e.g. layout, color), behavior (e.g. interaction, functionality) and structure (e.g. pagination, sections). Significant properties must be preserved over time for the digital object to remain accessible and meaningful.⁹¹

SIP (Submission Information Package): The data and metadata received into an OAIS-system at Accessioning as part of the ingest process. See also: AIP, DIP.⁹²

Stabilization: Establishing a safe and secure digital environment for the long-term preservation and storage of electronic records.

TAPER (Tufts Accessioning Program for Electronic Records): A tool developed by Tufts University to create submission agreements for electronic records. Flexible enough to apply to many types of born-digital materials.⁹³

TDR (Trusted Digital Repository): A standard developed by RLG and OCLC to define the characteristics of a sustainable digital archive that could serve large-scale, heterogeneous collections held by such research repositories as national libraries, university libraries, special collections, archives, and museums.⁹⁴

⁸⁸ Source: <http://www.archivists.org/news/custardproject.asp?prnt=y>

⁸⁹ Source: <http://rubyonrails.org>

⁹⁰ Source Roe pg. 61

⁹¹ Source: <http://www.jisc.ac.uk/whatwedo/programmes/preservation/2008sigprops.aspx>

⁹² Source: CCSDS Recommendation for an OAIS Reference Model, pg 1-13

⁹³ Source: <http://sites.tufts.edu/dca/about-us/research-initiatives/taper-tufts-accessioning-program-for-electronic-records>

⁹⁴ Source: <http://www.oclc.org/research/activities/past/rlg/trustedrep/>

TRAC (Trusted Repository Audit and Certification): A set of criteria to facilitate the certification of digital repositories capable of reliably storing, migrating, and providing access to digital collections. Developed by RLG and the US National Archives and Records Administration of the United States.⁹⁵

Virtual machine: A "completely isolated guest operating system installation within a normal host operating system". It is a software implementation of a machine (i.e. a computer) that executes programs like a physical machine.⁹⁶

⁹⁵ Source: <http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying-0>

⁹⁶ Source: <http://www.griffincaprio.com/blog/2006/08/virtual-machines-virtualization-vs-emulation.html>

Appendix B: Bibliography

- AIMS. (2011, October). *Born-Digital Archives*. Retrieved from: <http://born-digital-archives.blogspot.com/>.
- Archivematica. (2011, October). *Archivematica Open Archival Information System*. Retrieved from: http://archivematica.org/wiki/index.php?title=Main_Page.
- ArchivesSpace. (2011, October). *ArchivesSpace: Next Generation Archival Description*. Retrieved from: <http://www.archivesspace.org/>.
- BitCurator. (2011, October). *BitCurator: Tools for digital forensics methods and workflows in real-world collecting institutions*. Retrieved from: <http://bitcurator.net/aboutbc/>.
- Bodleian Electronic Archives and Manuscripts. (2011, October). *futureArch*. Retrieved from: <http://www.bodleian.ox.ac.uk/beam/projects/futurearch>.
- California Digital Library. (2011, October). *California Digital Library*. Retrieved from: <http://www.cdlib.org/>.
- Carolina Digital Repository Blog. (2011, July). About the Curator's Workbench. Retrieved from: <http://www.lib.unc.edu/blogs/cdr/index.php/about-the-curators-workbench/>.
- Carroll, L. L. (2010, October). The Salman Rushdie Papers at Emory University: Processing a born digital manuscript collection. In Personal Archives Institute. Conducted by Academy of Canadian Archivists, Toronto, ON.
- Charlesworth, Andrew. (2009). *Digital Lives: Legal & Ethical Issues*. London: British Library. Retrieved from: <http://britishlibrary.typepad.co.uk/files/digital-lives-legal-ethical.pdf>
- City of Vancouver Digital Archives. (2010). City of Vancouver Digital Archives system workflow v.1. Retrieved from: http://artefactual.com/wiki/images/4/40/COV_Digital_Archives_System_Workflow_v1.pdf.
- 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>
- Data Seal of Approval Board. (2010). Quality guidelines for digital research data. Retrieved from: http://www.datasealofapproval.org/sites/default/files/DSA%20booklet_2-0_engels_mei2010.pdf.
- Dooley, J. & Luce, K. (2010). Taking our pulse: The OCLC Research survey of special collections and archives. Dublin, OH: OCLC Research. Retrieved from: <http://www.oclc.org/research/publications/library/2010/2010-11.pdf>.

- DuraSpace. (2011). *Hypatia*. Retrieved from the DuraSpace Wiki: <https://wiki.duraspace.org/display/HYPAT/Home>.
- Emory Libraries. (2011, October). *Salman Rushdie's Digital Life*. Retrieved from MARBL: Manuscript, Archives & Rare Book Library website: <http://marbl.library.emory.edu/innovations/salman-rushdie>.
- Erway, R. et.al. (2011). Well-intentioned practice for putting digitized collections of unpublished materials online. Dublin, OH: OCLC Research. Retrieved from: <http://www.oclc.org/research/activities/rights/practice.pdf>.
- Forensic Investigation of Digital Objects (FIDO). (2011, October). Retrieved from: <http://fido.cerch.kcl.ac.uk/>.
- Forstrom, M. (2009). Managing electronic records in manuscript collections: A case study from the Beinecke Rare Book and Manuscript Library. *American Archivist*, 72(2), 460-477.
- Goethals, A. & Gogel, W. (2010, September). Reshaping the Repository: The Challenge of Email Archiving. In 7th International Conference on Preservation of Digital Objects (iPres), Vienna, Austria. Retrieved from <http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/goethals-08.pdf>
- Google. (2011, October). Snippet View. Retrieved from the Google Books website: <http://books.google.com/googlebooks/screenshots.html#snippetview>.
- Greene, M. A., & Meissner, D. (2005). More product, less process: Revamping traditional archival processing. *American Archivist*, 68(2), 208-263.
- Greene, M. A. (2010). MPLP: It's not just for processing anymore. *American Archivist*, 73(1), 175-203.
- Gueguen, G. et.al. (2011, October). *Day of Digital Archives*. Retrieved from <http://dayofdigitalarchives.blogspot.com/>.
- Hangal, S., Lam, M. & Heer, J. (2011). MUSE: Reviving memories using email archives. In *Proceedings of the 24th ACM Symposium on User Interface Software and Technology (UIST)* Santa Barbara, CA. Retrieved from: <http://mobisocial.stanford.edu/papers/uist11m.pdf>.
- Harvey, Ross. (2010). *Digital Curation: A How-to-do-it Manual*. New York: Neal-Schuman Publishers.
- Haynes, J. & Thompson, D. (2009, May). Accession to Access: Born Digital Archives in the Wellcome Library. In Future ProofV conducted by the International Scientific Archives Conference, Barcelona, Spain. Retrieve from: <http://files.me.com/fxroque/8swmac>.
- Hilton, C., & Thompson, D. (2007). Collecting born-digital archives at the Wellcome Library. *Ariadne*, 50 (January). Retrieved from <http://www.ariadne.ac.uk/issue50/hilton-thompson/>.
- Hilton, C., & Thompson, D. (2007). Further experiences in collecting born-digital archives at the Wellcome Library. *Ariadne*, 53(October). Retrieved from <http://www.ariadne.ac.uk/issue53/hilton-thompson/>.

- InterPARES I Authenticity Task Force. (2002). Requirements for Assessing and Maintaining Authenticity of Electronic Records. Vancouver, BC: The InterPARES Project. Retrieved from: http://www.interpares.org/display_file.cfm?doc=ip1_authenticity_requirements.pdf.
- Hobbs, C. (2007). The character of personal archives: Reflections on the value of records of individuals. *Archivaria*, 52(Fall), 126-135.
- John, J. L. (2008, September). Adapting existing technologies for digitally archiving personal lives: Digital forensics, ancestral computing, and evolutionary perspectives and tools, In Fifth International Conference on Preservation of Digital Objects, London, UK. Retrieved from: http://www.bl.uk/ipres2008/presentations_day1/09_John.pdf.
- John, J. L. et al. (2009). *Digital Lives – Personal Digital Archives for the 21st Century: An Initial Synthesis*. London: The British Library. Retrieved from: <http://britishlibrary.typepad.co.uk/files/digital-lives-synthesis02-1.pdf>.
- Kirschenbaum, M. et al. (2009). *Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use*. White Paper to the NEH Office of Digital Humanities. Washington, D.C. Retrieved from: <http://www.neh.gov/ODH/Default.aspx?tabid=111&id=37>.
- Kirschenbaum, M. et al. (2009, October). *Digital materiality: Preserving access to computers as complete environments*. In The Sixth International Conference on Preservation of Digital Objects. Proceedings (iPres), San Francisco, California. Retrieved from: <http://escholarship.org/uc/item/7d3465vg>.
- Kirschenbaum, M., Ovenden, R., & Redwine, G. (2010). *Digital forensics and born-digital content in cultural heritage collections*. Washington, D. C.: Council on Library and Information Resources. Retrieved from: <http://www.clir.org/pubs/reports/pub149/pub149.pdf>.
- MacNeil, H. (1994). Archival theory and practice: Between two paradigms. *Archivaria*, 37:10, 6-20.
- 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.
- Open Planets Foundation. (October, 2011). Retrieved from: <http://www.openplanetsfoundation.org/projects>.
- Orange Grove Digital Repository (OG). (October, 2011). *The Orange Grove: Florida's Digital Repository*. Retrieved from: <http://www.theorangegrove.org/about.asp>.
- Padfield, T. (2007). *Copyright for archivists and records managers*, third edition. London: Emerald Group Publishing Limited.
- Paradigm Project. (2007). *Workbook of digital private papers*. Retrieved from <http://www.paradigm.ac.uk/workbook>

- Pearce-Moses, Richard, ed. (2005). *A Glossary of Archival and Records Terminology*. Chicago: The Society of American Archivists. Retrieved from: <http://www.archivists.org/glossary/>
- Prom, Chris. (2011, October). *Practical E-Records*. Retrieved from: <http://e-records.chrisprom.com/>.
- Research Libraries Group (RLG) & National Archives and Records Administration (NARA). (2007). *Trustworthy Repositories Audit & Certification: Criteria and Checklist*. Retrieved from: http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf
- Roe, K. (2005). *Arranging & describing archives & manuscripts*. Chicago: The Society of American Archivists.
- Stollar Peters, C. (2006). When not all papers are paper: A case study in digital archivy. *Provenance*, XXIV, 23-35.
- Thomas, S., & Martin, J. (2006). Using the papers of contemporary british politicians as a testbed for the preservation of digital personal archives. *Journal of the Society of Archivists*, 27(1), 29-56.
- Tufts University Digital Collections and Archives. (2011, August 9). TAPER: Tufts Accessioning Program for Electronic Records. Retrieved from Tufts website: <http://sites.tufts.edu/dca/about-us/research-initiatives/taper-tufts-accessioning-program-for-electronic-records/>
- Tufts University Digital Collections and Archives and Yale University Manuscripts & Archives. (2006). *Fedora and the preservation of university records project: 3.1 maintain guide, version 1.0*. Retrieved from: <http://hdl.handle.net/10427/1286>.
- University Of Michigan. (2011, September 29). Preservation and format support. Retrieved from Deep Blue website: <http://deepblue.lib.umich.edu/about/deepbluepreservation.jsp>.
- Wellcome Library. (2011, September 29). Digital curation toolbox. website: <http://library.wellcome.ac.uk/node289.html>.
- Zhang, H. & Twidale, M. (2010). *The folders we live in: What we need and what we can get*. Technical Report, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. Retrieved from: <http://hdl.handle.net/2142/16610>
- Zhang, J. (2010). *The Principle of Original Order & the Organization and Representation of Digital Archives*. Dissertation. Graduate School of Library and Information Science, Simmons College.

Appendix C: Contributor Biographies

Nicole Bouché

Director of the Albert and Shirley Small Special Collections Library, at the University of Virginia Library

Nicole arrived at UVa shortly after the AIMS Project was launched. She provided general oversight and guidance to the project team at Virginia. Previously, as Manuscript Unit Head at the Beinecke Rare Book and Manuscript Library from 1993-2005, she established Beinecke's earliest policies and procedures for preserving, documenting, and providing patron access to born-digital content that was being acquired by Beinecke amongst collections of personal papers and literary manuscripts. She is active in the Rare Book and Manuscript Section of ACRL/ALA, as well as other professional associations.

Judy Burg

University Archivist, University of Hull at Hull History Centre

As University Archivist, Judy manages holdings of around 2800 linear metres (9200 linear feet) and a core team of two staff, augmented by fixed-term project staff (such as the AIMS Digital Archivist). With this small core team, her role includes involvement in most activities relating to archive stewardship, with an emphasis on collection development and strategic planning. She also undertakes the teaching of archive information literacy and research skills, and the development of new teaching, research and public engagement activities in collaboration with academics, particularly within arts and humanities. She is also a member of the strategic leadership team within the Directorate of Library & Learning Innovation.

Hull University Archives is based at Hull History Centre, a joint facility and service operated in partnership between the University and Hull City Council. The Centre encompasses storage, management, conservation, access, outreach and education functions relating to archives and local studies for its two parent bodies. Ownership and/or custodianship of archives remain separate within each organization, but all other functions are undertaken jointly. Judy's role therefore also involves joint strategic leadership for the Centre, in collaboration with the City Archivist and she was joint lead for the project to establish the Centre 2004-2009. The Centre has total archive and special collections holdings of over 8,000 linear meters (26,000 linear feet) and a team of 15.5 FTE staff.

Prior to the establishment of the AIMS project, born-digital collections held by Hull University Archives were minimal (less than 1 MB, held on legacy media). Archives were however included as a use case during the development of the University's digital repository, 2005-2008 and Judy was involved in projects investigating process requirements for records management and archives collection development.

Before taking up the role of University Archivist at Hull in 2003, Judy was Company Archivist for The Boots Company PLC (now Alliance Boots). She is a member of the Political Parties and Parliamentary Archives Group UK (PPPAG) and Secretary of GLAM (Group for Literary Archives and Manuscripts).

Peter Chan

Digital Archivist, Stanford University

Peter is primarily responsible for setting up and managing the workflow for pre-accessioning, accessioning, processing, delivering and preserving of born-digital archival materials acquired by the Special Collections Department at the Stanford University Libraries. The Digital Archivist position hired with grant funds was made permanent as a split between Special Collections and DLSS cementing the collaborative efforts for born-digital processes. Peter will continue to serve in this capacity at the close of the grant-funded period.

Peter has explored several new techniques and workflows to accommodate born-digital materials at Stanford. Peter explored the use of "High Resolution" photograph as an enhanced curation work. He also lead the development of the AIMS Digital Material Survey to guide archivists/cruators in the discussion of collecting digital material from donors. Peter also explored different hardware to connect a 5.25 floppy drive to a modern personal computer and built a capture station with a 5.25 floppy drive, a ZIP drive from an old computer, new motherboard, hard drives, RAM, video card, power supply and computer case bought from an electronic store. On the software side, Peter evaluated FTK Imager, Tableau Imager, AccessData FTK, and Karen's Directory Printing for the accessioning task. Peter also piloted the use of AccessData FTK, a forensic software, in extracting technical metadata, assigning rights and descriptive metadata to born digital materials. He evaluated the use of Emailchemy to convert closed and proprietary email formats to standard formats based on RFC-2822 and Transit Solution to convert files in more than 200 formats to HTML files as display derivatives. Peter also collaborated with Stanford scholars on novel delivery options for email.

Prior to joining the AIMS project, Peter was an assistant archivist at Stanford University Libraries. He processed the born-digital materials from the Richard Fikes papers, SUMEX (Stanford University Medical Experimental Computer 1974-1983) collection, John McCarthy's web site, and Nils Nilsson papers.

Bradley Daigle

Director of Digital Curation Services, University of Virginia Library

Bradley Daigle is Director of Digital Curation Services and Digital Strategist for Special Collections at the University of Virginia Library. He is also co-PI on a Andrew W. Mellon Foundation grant entitled “Born Digital Materials: An Inter-Institutional Model for Stewardship (AIMS).” Having been in the library profession for over a decade, he has published and presented on a wide range of topics including mass digitization, digital curation and stewardship, sustaining digital scholarship, intellectual property issues, and mentoring in libraries. His research interests include the history of the book, early modern natural history, and James Boswell. He received his MA in literature from the University of Montreal and an MLS from Catholic University.

Glynn Edwards

Principal Manuscripts Processing Librarian, Stanford University Libraries & Academic Information Resources

As the head of the manuscripts unit in Special Collections, Glynn manages holdings of more than 38,000 linear feet of manuscripts material in all formats and supports the efforts of over ten subject curators, primarily in the Humanities, Social Sciences and Area Studies groups.

Due to the large backlog of born digital media — over 25,000 legacy items — Special Collections and the Digital Systems and Services Department (DLSS) at Stanford entered into a partnership in early 2009. Working together they designed phase one of a forensic recovery lab to develop sustainable and long term strategies for capturing data stored on legacy media. In conjunction with this effort, Glynn joined the Stanford team on the AIMS project and began a two-year process to flesh out strategies for accessioning, processing and delivering born-digital content.

Glynn is also involved on campus in the Digital Initiatives Group Plus which serves to connect those working with digital collections and in the humanities at SULAIR. She is part of the sub-group responsible for holding monthly discussions on various digital topics. The AIMS team at Stanford is now in the process of expanding into a Born Digital Working Group to include the University Archivist, archivists and librarians from other repositories on campus, etc. Their main goals are to address issues that were not tackled during the course of the grant, develop policies and guidelines at Stanford, and refine and continue to develop and expand strategies and workflows, and to develop a robust internship program to train local library students in all aspects of accessioning and processing of born digital material. The working group is also involved in plans for phase two of the digital forensic lab.

Glynn was previously the Associate Head of Collection Services, Manuscripts at the Schlesinger Library at Radcliffe Institute, Harvard University. There she was responsible for the creation, validation, and export of EAD guides; registration of digital objects in HUL's Name Repository Services, creation of links in EAD to digital objects and depositing digital objects into HUL's Digital Repository, and maintenance of the library website. She was also a member of their OASIS group – which managed the platform for delivering finding aids at Harvard.

Michael Forstrom

Archivist, Beinecke Rare Book & Manuscript Library, Yale University Library

Michael's chief responsibilities at the Beinecke include processing and cataloguing of literary archives and materials and stewardship of born-digital archival material. He has been responsible for stewardship of born-digital archival material since 2002. As such he collaborates with library management, curators, information technology staff, archivists, university offices, and professional communities to establish best practices for managing born-digital material acquired by the library.

Michael was formerly archivist at the National Opinion Research Center (NORC) at the University of Chicago, 2001-2002. He is a member of SAA, GLAM-NA, and MSA

Kevin Glick

Head of Digital Information Systems and University Archives, Yale University Library

Kevin is the Head of University Archives and Electronic Records Archivist at Yale University's Manuscripts and Archives <http://www.library.yale.edu/mssa/>. Since the beginning of 2009, major work projects have included completing the migration of the department's legacy collection management data and descriptive finding aids for over 100,000 boxes and 2700 collections into the Archivists' Toolkit <http://www.archiviststoolkit.org/>. Kevin is an adjunct faculty member at Southern Connecticut State University and Simmons College. Prior to joining the Yale staff in 2002, he was the project manager of the US team of the InterPARES Project <http://www.interpares.org/>. Kevin holds an M.L.S. from the University at Albany, SUNY; an M.A. in medieval studies from Western Michigan University; and a B.A. in history from Ohio.

Gretchen Gueguen

Digital Archivist, University of Virginia Library

As a member of the Albert and Shirley Small Special Collections Library, Gretchen is primarily responsible for all born-digital materials acquired by the Albert and Shirley Small Special Collections Library. Her role is an evolving one within the library and includes collaboration with colleagues involves in all aspects of collection management. Gretchen was preceded as Digital Archivist for the AIMS project by Elizabeth Gushee.

Prior to joining the Library at Virginia in May of 2011, Gretchen was head of Digital Collections at East Carolina University's Joyner Library, responsible for digital repository services and digitization. These projects included digital content management systems for both digitized and born-digital materials from the special collections and archives, born-digital student works in multiple formats as well as institutional electronic records. She has been involved in digital library and digital humanities projects since earning her MLS at the University of Maryland in 2005.

Gretchen has focused on project management within special collections and digital library projects throughout her career. She has been involved in the Society of American Archivists, the Library Information Technology Association of the American Libraries' Association as well as the Digital Humanities conference, and several state wide digitization projects within the state of North Carolina.

Tom Laudeman

Project Software Engineer, University of Virginia Library

As the developer for the AIMS project Tom's role is to provide technical support to the AIMS Digital Archivists. This includes technical assessments of existing software as well as writing new packages, especially Rubymatica. In order to illustrate extant issues in digital archive practice, Tom has written several blog entries at the AIMS blog.

Tom's career could be characterized as creating user interfaces to data. In the beginning is the data, then come tools to visualize and manipulate that data. His work has been as varied as online gaming, bioinformatics, genomics, business process automation and the field once known as "desktop publishing". Working primarily on Linux systems and open source software, for the last dozen years Tom's projects have mostly been browser agnostic, dynamic web sites with SQL data stores. Tom is a content creator and has written several content management systems to manage his 1400 images and 1300 web pages.

Mark Matienzo

Digital Archivist, Manuscripts and Archives, Yale University Library

Mark was seconded onto the AIMS project immediately after starting at Yale in January 2010 as a digital archivist. At Yale, Mark is responsible for developing and implementing workflows and procedures for handling electronic records, and I support many of the technology requirements of my department, including our Archivists' Toolkit deployment. He also serves or has served in an advisory capacity on a number of committees, including that which oversees development of the Yale Finding Aid Database, as well as a campus-wide digital preservation committee. I also occasionally provide service as a reference archivist.

Currently his focus has been shifted towards improving Yale's workflow for accessioning of electronic records, which we have been building collaboratively with the Beinecke Rare Book and Manuscript Library.

Prior to the AIMS project Mark had a minor amount of experience with electronic records. However, he has had experience with Fedora in a previous position overseeing the technical aspects of a digital repository project at the New York Public Library.

Before joining Yale, Mark worked as an Applications Developer for Strategic Planning (formerly the Digital Experience Group) of the New York Public Library, where he had a wide variety of responsibilities, including back-

end web development for online exhibits and the Library's new Drupal-based website, and serving as the technical project manager for the Library's digital repository initiative. He has also worked as an assistant archivist at the Niels Bohr Library and Archives at the American Institute of Physics, as a project archivist at the Smithsonian Institution's National Anthropological Archives, and as a cataloger at ProQuest Information and Learning. He has also been a consultant for the Philadelphia Area Center for History of Science, the Brooklyn Historical Society, and the ArchivesSpace planning grant.

Mark has held a wide variety of leadership and service positions in the Society of American Archivists, focused mostly on descriptive practice and standards. He is currently Co-Chair of the SAA Encoded Archival Description Roundtable, a member of the Schema Development and Review Team of the SAA Standards Committee, and an ex officio member of the Technical Subcommittee for Encoded Archival Description and the Technical Subcommittee for EAC-CPF. He has also undertaken considerable research investigating the use of open source forensic software to support archival workflows.

Michael Olson

Digital Collections Project Mgr & Technologist, Special Collections, Stanford University Libraries & Academic Information Resources

Michael recently led the creation of SULAIR's Digital Forensics Lab to preserve and provide access to born digital collections. Michael has an M.Phil in History and Computing from the University of Glasgow, Scotland and a B.A. in Medieval Studies from the University of British Columbia, Canada.

Simon Wilson

Digital Archivist, Hull University Archives

Simon was seconded to the role of Digital Archivist from his post of Senior Archivist at the University. On taking-up this appointment in November 2008 his main priority was to prepare the collections for the move from the University library to the Hull History Centre. This new purpose-built centre brings together three separate institutions to provide a single point of access. Simon was also responsible for managing all ICT aspects within the new building including staff, public and wifi networks and the development of a dedicated website for the Centre and the creation of an integrated online catalogue.

As part of the small archives team Simon works across all aspects of the service from collection development, cataloguing and providing public access to the collections held at the History Centre as well as some teaching relating to research skills and highlighting potential collections within the archives that are relevant to particular under-graduate modules for students studying History, English and Politics.

After qualifying as an archivist in 1995 Simon has worked on a number of cataloguing projects across the higher education, local authority and charity sectors but had no previous experience in processing or managing born-digital archives. Prior to his post at the University of Hull Simon was Collections Project Manager at Hull Museums (2005-2008) on a retrospective-documentation project working across 7 museums including the selection and implementation of a collections management system. As Mersey Gateway Project Manager (2001-2004) he led a project that resulted in the digitization of over 20,000 items from archives, libraries and museums across the North West.

Simon is a registered member of the Archives and Records Association and a mentor to two recently qualified archivists on the Registration Scheme. He is currently Secretary to the Association's Data Standards Group. He is also part of a small working group looking at Digital Archives issues on behalf of CALM users.

Appendix D: Institutional Summaries and Collection Descriptions

I. The University of Hull, University Archives at Hull History Centre

Hull University Archives is part of the Research and Learning Resources Group (RLR) within the Directorate of Library and Learning Innovation (LLI). The remit of LLI includes all library services (acquisition, management and circulation) and also includes strategic management and development of the University's digital repository.

The University Archives and associated staff are based off-campus within Hull History Centre. This is a joint service, in partnership with Hull City Council, encompassing local studies library material and archives. The Centre itself has no independent legal identity, so ownership and custodianship of archives remain separate within each partner organization. However, all services, including access, remote enquiries, outreach and conservation are jointly operated.

The core University Archives team consists of three full-time posts:

- University Archivist
- Senior Archivist
- Archives Assistant (currently 0.5% FTE)
- Within the larger Hull History Centre team there are currently 15.5 FTEs, including 6 archivists and a conservator

The University has collected manuscripts and archives since the establishment of Hull University College in 1928 and holdings now run to around 2800 linear metres (c. 9200 linear feet), including 120 larger and over 300 smaller collections. The documents date from the late 11th to the early 21st centuries. Collecting specialisms in pressure groups and politics, and in modern literature and drama were established during the 1960s, under the influence respectively of Professor John Saville and the University Librarian, Philip Larkin. In addition to these two areas, and the archives of the University itself, the acquisition policy now also encompasses archives relating to maritime history.

The University's rare book collections remain on campus, managed and accessed within the Brynmor Jones Library, the sole library at the Hull campus. There is also a smaller library on the University's Scarborough campus.

At the outset of the AIMS project the University of Hull Archives only had isolated digital media amongst its collections, accessioned as physical objects only and not as digital material. This legacy material has been reviewed during the AIMS project and much has now been fully accessioned, enabling us to develop workflow and processes. As a result of participation in the AIMS Project and discussion with project partners the archives have now established a forensic workstation including an offline PC for virus checks and other accessioning processes with two Tableau write-blockers for use with a range of hard drives from PCs and laptops. We have also created supporting documentation, workflow and procedural guidelines to enable the archives to receive born-digital content from a range of media formats.

The main collections management tool is Axiell's DS CALM. This includes linked information about depositors/donors and accessions as well as hierarchical catalogue entries, at collection, series, in some cases sub-series, and item level. The location register, conservation logs and some legacy collections management data remain in MS Word and Excel files and in paper form.

For Hull History Centre users, a web-based version of the University's CALM catalogue, merged with CALM databases covering the City Archives and the Local Studies collections provides the main discovery route. It includes collection-level and item-level information. There is currently no direct integration with the University's library catalogue, although there is increasing integration within web-based source guides and associated information skills sessions for students. LLI is investigating the use of Blacklight as a common interface to internally held catalogues and collections, including collections held in the University's digital repository. (Both the repository and the use of Blacklight are described in more detail below.) Digital archive files (both born-digital and digitized) will be stored within the repository once processed and will be linked to the relevant catalogue information in CALM. Therefore the use of Blacklight offers the potential to provide a discovery and access route for both library and archive resources, both paper-based and digital, in one on-line location.

Alongside the library system and online catalogue LLI operates the institutional digital repository service, launched in 2008 and based on the Fedora digital repository system, which holds a variety of open and restricted access digital collections. These encompass teaching (e.g., open educational resources, exam papers), research (e.g., publications, datasets), and administration (e.g., committee papers, HR documents). The repository has also implemented the Hydra repository interface system to facilitate the presentation and management of different content types. The public face of the repository is provided via Blacklight (used as part of Hydra), and repository collections are also a component of the potential further use of Blacklight.

I. Papers of Stephen Gallagher.

Stephen Gallagher, Hull alumnus, is a novelist, screenwriter and director specializing in contemporary suspense. 42 boxes (7 linear metres, 23 linear feet) of paper records relating to his early works including *Doctor Who*, *Bugs* and *Chimera* were deposited in 2005. As part of his participation in the AIMS project he transferred some born-digital

records (14,320 files, 13.6GB) via an external hard drive. The born-digital material includes more recent work including *Eleventh Hour* and *Crusoe*, a previous version of his website and content from his blog (<http://brooligan.blogspot.com>). There was also a large number of saved webpages that reflect an element of his research process and some of his on-going work, including his latest novel *The Suicide Hour* (prior to publication) and a number of pilots sold to USTV networks in 2010.

The material is significant in the context of UK and US television drama of the past 20 years, and it also reflects and demonstrates the creative, promotional, and production processes associated with novel and screen writing, from concept to broadcast in the paper and then the born-digital environment.

The particular issues faced with this collection include the presence of 39 Amstrad disks and about 300 files created using specialist screenwriting software *FinalDraft*. The presence of over 80 webpages saved from the web with their associated files (1226 files, 14.5MB in total) also brought with it copyright and presentational issues.

The collection has been processed with a view to creating a single integrated catalogue to both the paper and born-digital components and once completed this will be made available via the History Centre online catalogue at <http://www.hullhistorycentre.org.uk/catalogue>.

In arranging the collection particular care was given to reflect the donor's methods of working (see *Arrangement and Description Case Study: The Papers of Stephen Gallagher*) whilst at the same time enabling easy discovery and access. We are currently liaising with the depositor about possible access restrictions to his most recent and currently unpublished work. Access to the born-digital material is likely to be via a locked-down PC in the Hull History Centre searchroom in the first instance. We are hoping to move towards a model of online access for some of the material in the next 2-3 years as part of the University's work using Hydra and Blacklight.

2. Socialist Health Association Papers

The Socialist Health Association (SHA) is a UK-based membership organization, affiliated with the Labour Party, which promotes health and well-being and the eradication of inequalities. A significant volume of paper (7 linear metres, 23 linear feet) material including minutes, reports, correspondence, circulars press releases, financial records, and photographs, dating back to 1930 had already been deposited with the archives. A tranche of born-digital material (2558 files, 670MB) was deposited by Martin Rathfelder, the Director of the SHA, as part of the AIMS project.

The particular issues faced with this collection were the possible integration of born-digital material into a pre-existing archival structure. Due to the way previous accessions had been catalogued discretely and not into a single system of arrangement complete integration was not possible so a distinct series has been created and it is hoped that this will be flexible enough to accommodate subsequent accruals of born-digital material.

There were no issues surrounding legacy media and the main content issues surround the presence of an Access database, the SHA website (1180 files in 51 folders, 86.4MB) and the 90 or so SHA e-mail newsletters that have been issued in the last two years. These three aspects all contain processing and presentational issues that need

further consideration before proceeding. By far the biggest concern is that relating to copyright, with a large number of presentations and other content having been produced by third-parties (e.g., Conferences folder contains 444 files in 21 sub-folders): it is presumed that online access is not appropriate for this material. There is also a need to appraise the material to remove blank forms and un-related material.

The collection has been processed and once completed this will be made available via the History Centre online catalogue at <http://www.hullhistorycentre.org.uk/catalogue>. Access to the born-digital material is likely to be via a locked-down PC in the Hull History Centre searchroom in the first instance. We are hoping to move towards a model of online access for some of the material in the next 2-3 years as part of the University's work using Hydra and Blacklight.

2. Stanford University, Stanford University Libraries & Academic Information Resources

The Digital Forensics Program at Stanford began as a collaboration between two units: the Digital Libraries Systems & Services Department (DLSS) and the Department of Special Collections & University Archives, Manuscripts (SPEC). For over 15 years the Manuscripts Division in SPEC has been recording gross extents of computer media in their accessions. This legacy media in our backlog reached 25,000 items in the winter of 2009 and initiated our build out of a forensic recovery lab and our subsequent involvement in the AIMS grant project.

At this stage, the Department of Special Collections & University Archives currently holds over 50,000 linear feet of materials (or over 80 million pages). The Manuscripts Division comprises 75% of the holdings and takes in an average of 1,800 linear feet per year. These holdings include over 28,000 items of computer media, 13,680 audiotapes, 10,416 videotapes/film and over 296,000 still images.

As we began the AIMS project in the fall of 2009, our project team consisted of Michael Olson, a project manager from the DLSS group, and Glynn Edwards, head of the Manuscripts Unit in SPEC. Tom Cramer, head of DLSS was Stanford's site manager. Three months into the project we hired Peter Chan as our Digital Archivist. Prior to the start of the AIMS project, we received our forensic equipment and began setting up our first forensic recovery lab. It was largely built around a Forensic Recovery Evidence Device - or FRED. That fall, staff from DLSS, University Archives, the Manuscripts Unit and our curatorial group attended training at Digital Intelligence; in February 2010, Peter attended more in-depth training in forensic toolkit software (FTK) used in working on case files. The following two years have been an intensive period of testing - both capture from legacy media and processing of born-digital materials using forensic tools.

Currently the Digital Forensics Program Working Group consists of Michael Olson (DLSS), Glynn Edwards (SPEC), Peter Chan (reporting jointly to DLSS/Spec) and Henry Lowood (Curator for the History of Science and Technology). Our program also began a series of open meetings with library staff from other departments and repositories on campus in the fall of 2011.

Our Digital Archivist, Peter Chan, runs the Digital Forensic Lab and assists curators and donors with any issues arising with new accessions. This is a base-funded position that reports to both managers – DLSS and SPEC. Peter also works closely with developers in DLSS in scripting out digital objects and metadata from new tools – like Forensic Toolkit and PhotoMechanic.

Staffing in the Manuscripts Unit – which co-manages the Digital Forensic Lab - is relatively light consisting of two full-time employees, including the head of the division, and 2 half-time employees. Our division uses Archivists Toolkit for both collections management and the creation of finding guides. These are exported to the Online Archive of California – a regional site – and aggregated on Archive Grid. Our collection-level catalog records are created using Sirsi Dynix and exported to OCLC.

Special Collections and the Digital Forensics Program will be conducting a pilot project in 2012 delivering processed collections of email – specifically from the Robert Creeley and Peter Koch collections (AIMS project) and possibly Stephen Schneider (processed by University Archives staff) – via our reading room. We are planning to present the email archives with an interface created by Sudheendra Hangal, a graduate student in Stanford's Computer Science Department, to facilitate browsing and conduct user tests to help direct future development.

I. Stephen Jay Gould Collection

Influential American paleontologist, evolutionary biologist and historian of science, Stephen Jay Gould began his career at Harvard University in 1967 where he worked until his death in 2002. One of the most popular science writers of our time, he is the author of 22 books, 479 peer-reviewed scholarly papers, 300 essays and 101 reviews.

At the time of the AIMS grant, the Gould collection consisted of eight accessions acquired between 2004 and 2010. Totalling over 500 linear feet of material, the collection contains writings, correspondence, research, juvenilia, specimens and legacy computer media. The papers and specimens were processed concurrently with the AIMS project.

Media enumerated initially consisted of: 60 5.25-inch floppy diskettes, 81 3.5-inch floppy diskettes, two cartons of computer punch cards and 3 computer tapes. The diskettes contain bibliographic databases and working drafts of many of Gould's publications. The punch cards and the data tapes appear to contain datasets used in his evolutionary biology research. Since the beginning of the AIMS project, we have uncovered more computer media (21 more sets of computer punch cards) in a later accession and odds and ends scattered within folders throughout the accessions.

Gould was the first collection we worked with and thus underwent several trial efforts both in capture and processing. The first attempts at capture created disk images using ImageTool™ and a Catweasel in FRED.⁹⁷ However, ImageTool™ did not generate either an audit log file to confirm successful imaging or a file listing of the disk contents. The second attempt was more successful and utilized an old personal computer with on-board

⁹⁷ A Catweasel is just an interface card for computer that does not have a floppy interface in the motherboard. Write-blocking is enabled by putting a tape at the "write-protect" area in a 5.25 inch floppy disk. FRED = Forensic Recovery Evidence Device.

floppy disk controller was used to image the diskettes using free software called FTK Imager™. Outputs from FTK Imager™ include: disk images, audit log files to confirm successful imaging and file listings of the diskette contents. Unreadable media – primarily a result of physical damage before transfer to SULAIR – was slightly over 6%.

The first efforts in processing – before we settled on FTK™ – used Windows Explorer to arrange the files and Quickview Plus™ to view their content. Folders were created that mirrored “series” and “subseries” in the concurrent processing project and files were moved from their original media folder into this new hierarchy. But this changed the original metadata associated with the files – such as original file path, etc. By this time, Peter had tested Forensic Toolkit (FTK). FTK extracted the technical metadata (file size, creation, last modification and last accessed dates, file format, checksum, etc.) of the files in the disk images loaded. “File Category” provided a summary of how many files are in different file formats. The interface to hide the duplicate files was activated so that users are working on unique files (FTK uses the checksums of the files to identify duplicate files).

Restricted content such as credit cards, social security number, student grades, etc. were identified using the pattern & full-text searches functions. The files identified were flagged as “Privileged” and will not be delivered to the public. Although the search may not find all the restricted contents, it allowed us to perform a good faith effort to do so that will be scalable moving forward.

Bookmarks were created with keywords that mirrored series and subseries titles in EAD for the papers. The embedded viewer (reads over 200 file formats) was used to view files during processing with obsolete file formats. Files were then assigned to bookmarks according to intellectual contents individually or in batches. The “Label” functionality in FTK was used to represent other crucial metadata, such as: access restrictions, document types, computer media type, and subject headings. Reports in XML/HTML format are generated to export files to access repository (Hypatia). The files carried the bookmarks, labels, privileged flag, and technical metadata with them.

All the material in the Gould collection will be described in the online finding aid, although the digital files will be described at the series level only. Notes regarding processing and capture methodology will be included here. There will be links in the final guide and the collection level catalog record to the digital contents in Hypatia. The files will be full-text searchable and delivered via the web, open to all (except those flagged as privileged).

2. The Papers of Robert Creeley

Robert Creeley is an American poet, novelist, short story writer, editor and essayist. Author of more than 60 books, Creeley taught at Black Mountain College (BMC) in the 1950s and was one of the Black Mountain poets, an avant-garde group of poets centered on BMC.

The Creeley collection comprises over 450 linear feet of materials with the last 100 feet of accessions received still unprocessed. The processed papers feature Creeley's own working manuscripts for his poems and critical writing, both published and unpublished. These appear in a variety of formats: notebooks, filled with autograph drafts of poems; typescripts, often annotated in holograph; frequent pieces written on random scraps of papers, as well as

over 50 items of legacy media containing files for individual poems and works of prose as well as email backups. The material on the 53 3.5" floppy diskettes, 5 Zip Disks, and 3 CD-ROMs was never captured and remained closed to researchers until included in the AIMS project.

After discussions with curatorial staff who stated that Creeley deleted files before transfer to Stanford on purpose, we decided to capture logical images rather than disk images in this instance. We used a floppy drive capture station, designed and built by Peter Chan, and AccessData's Imager software. There were some issues that hampered our efforts. The first were backup files and proprietary software on the Zip disks. Five of the six disks contained backup files unrecognizable by Forensic Toolkit (FTK) – the software we decided to use for processing these materials. These backup files on two of the disks were possibly created using the proprietary backup software originating from Iomega (the company which made the Zip disks); the files on one recognizable and were likely copied using Windows Explorer.

Another issue was that the number of files gave us a bit of a challenge in ascertaining how many files there actually were! First, some files were zipped on the computer before copying to floppy diskettes and CDs. And, some emails were copied as one file per email and others in the "MBOX" format which contained thousands of emails in one MBOX file. After processing, it appeared that there were approximately 50,000 original emails rather than the initial estimate of 80,000.

Our intent is to describe the digital content at the series level and incorporate it into the existing finding aid online. The digital content will be delivered in two ways. Creeley's writings will be delivered via Hypatia (end of October release) while email, because of the multiple recipients and senders, will be delivered via a stand-alone computer in the Reading Room. In order to extract some useful information from the emails for indexing purposes, we tested the use of network diagrams.

The header information ("to", "from", "subject" and "date" fields) for 50,000 unique emails were output as a *.csv file using a utility in FTK. A Digital Humanities expert at Stanford University Libraries, Elijah Meeks, opened the file in Gephi⁹⁸, open-source software for visualizing and analyzing large networks graphs, to create network diagrams. These diagrams show the names of correspondents as well as the movement of correspondence between authors and recipients.⁹⁹

To conclude, in 2011 we received another 25 feet of Creeley material, which has not been processed as part of the project. It contained the following computer media: 7 computers (Compaq Presario CQ60 Notebook PC with Windows 7 [owned by the dealer]; SONY PCG-321A Notebook PC with Windows ME; SHARP Actius MM20 Notebook PC with Windows XP; Gateway Solo Notebook PC; Dell MTC2 Desktop PC; Midwest Micro Desktop PC; Racer Desktop PC); 3 zip drives; 121 optical discs; 422 3.5-inch floppy diskettes; 1 Olympus Camedia CF/

⁹⁸ <http://gephi.org/>

⁹⁹ Elijah published an article on the Digital Humanities site at SULAIR - <https://dhs.stanford.edu/visualization/robert-creeley-e-mail-correspondence-network/>

SmartMedia Reader; 1 Zip 250 USB Drive; 1 Olympus C-4000 Camedia Digital Camera; 1 8-megabyte Olympus SmartMedia Flash memory card; 1 128-megabyte SanDisk SmartMedia Flash memory card; 1 20-gigabyte iPod.

The dealer informed us that he had transferred contents of all of Creeley's computers, Zip disks, and CD-ROMs as well as some of the floppies to the new Compaq laptop computer. He also mentioned that some media contained files that appeared corrupt or were unable to be copied.

3. Peter Koch Collection

Contains one hard drive with correspondence and graphic arts files. Black Stone Press ephemera, 1974-1995. Peter Koch got his start in printing in Missoula, Montana when he founded the Black Stone Press, a publishing imprint and letterpress printing office, in tandem with artist Shelley Hoyt, in 1974. Four years later, the press relocated to San Francisco. Koch has operated his own design and printing studio continuously for almost thirty years. A creative force and personality in Bay Area fine press book design, printing, and publishing, Koch's work has earned an international reputation. His works include editions of ancient Greek philosophers, the musings of maverick poets, and the images of world-renowned wood engravers and photographers. Koch specializes in publishing limited edition livres d'artistes, broadsides, portfolios, and what Koch describes as —text transmission objects. Koch is the co-founder of the CODEX Foundation, a non-profit organization devoted to promoting and preserving the arts of the book. The image files (RAW, TIFF, JPEG, etc.) in this collection will be searchable and normalized for delivery out of the Hydra client and/or Stanford's digital collections delivery portal (<http://collections.stanford.edu>). The original Quark (design files) will be delivered via download in their original binary format to interested researchers. Overall description of projects and files on the hard drive would be listed as a separate series, echoing the existing arrangement to some degree of the physical collection (currently in Archivists' Toolkit). The collection also includes email that will need to be vetted with the donor for accessibility.

4. Xanadu Project collection.

Contains 6 hard drives with papers relating to the Xanadu Project, XOC, and Eric Drexler. The Xanadu Project was founded in 1960 by Ted Nelson, and was the first hypertext project, widely regarded as a conceptual antecedent of today's World Wide Web. The contents of the hard drives in this collection will be described in Archivists' Toolkit and linked to from the finding aid. Selected files of interest will be made directly accessible via the Hydra client and/or Stanford's digital collections portal; disk images of the entire hard drive will also be made for preservation purposes, and may be made accessible (based on the judgment of the archivist).

3. The University of Virginia, Albert and Shirley Small Special Collections Library

The Albert and Shirley Small Special Collections Library administers over 13 million manuscripts, 3.6 million items in the University archives, and 325,000 rare books, as well as approximately 3,000 maps, over 4,000 broadsides; more than 250,000 photographs and small prints; over 8,000 reels of microfilm; and substantial holdings of audio recordings, motion picture films, printed ephemera, and a growing number of born-digital resources which, to date, arrive chiefly as components of contemporary archival and manuscript collections. The Library occupies a new building on the University's historic Grounds, which features state-of-the-art climate control and security for the University's special collections, a new reading room, a seminar classroom and auditorium, and permanent and changing exhibitions in two galleries.

The Library is perhaps best known for its extensive collections, both printed and manuscript, related to American history and literature. Highlights include Virginia; papers relating to Thomas Jefferson, his family and descendants; the *Albert H. Small Declaration of Independence Collection*; rare books and maps related to early European voyages of discovery and exploration, especially in North America, in the *Tracy W. McGregor Library of American History*; and sources, both printed and manuscript, relating to African-American history, particularly in Virginia and the South. The *Clifton Waller Barrett Library of American Literature* forms the cornerstone of the American literature collections, supplemented by other substantial literary holdings. Other collection highlights include the *Joseph M. Brucoli Great War Collection*, the *Douglas H. Gordon Collection* of early French books and fine bindings, the *Paul Victorius Evolution Collection*, the *Marion duPont Scott Sporting Collection*, the *McGehee Miniature Book Collection*, special topics in British literature, including the *Sadleir-Black Gothic Novel Collection*, holdings from the Paul Mellon library; extensive collections of American and European sheet music and scores; the *Martin Jules Hertz Collection of Classical Pamphlets*; the *Franz Kafka Collection*; the *Wilbur Cortez Abbott Collection of Seventeenth-Century English History and Literature*; and the *Jorge Luis Borges Collection*. The Library also holds noteworthy material and collections in the history of books, typography, and printing, spanning the period from the very earliest printers' manuals to those of the present day, and including the productions of fine private presses and contemporary artists' books.

The Small Library also houses the University of Virginia Archives, documenting the history of the University since its founding by Thomas Jefferson, including many of Jefferson's original architectural drawings and notes for his "Academical Village", which is now a UNESCO World Heritage Site.

The Small Library employs 18 FTE, including 4 staff whose activities are devoted largely, but not exclusively, to management of the Library's archival and manuscript collections.

The UVa Library employs several teams to support the management of its digital assets. In total, there are 14 staff directly related to such activities, though the number is far greater for all the stages involved in the stewardship of born digital materials. For our technology stack, we employ the Hydra Stack (see <https://wiki.duraspace.org/display/>

[hydra/The+Hydra+Project](#)). We also have Quantum's StorNext HSM software for the backup and preservation of our archival masters (http://en.wikipedia.org/wiki/StorNext_File_System).

The Library is actively engaged in efforts to scan its most rare and unique out of copyright holdings, to make them web-accessible, worldwide. As a result of the AIMS project, a full-time Digital Archivist has joined the staff and is developing workflows for the accession, processing, discovery, and management of born-digital materials. A Forensic Recovery of Evidence Device (FRED) (see <http://www.digitalintelligence.com/products/fred/index.php>) has been purchased along with accompanying drives for various hardware formats.

1. Alan Cheuse papers, 1976-2009.

The files of author, book critic, and NPR's "voice of books" Alan Cheuse whose collection includes electronic drafts of novels and short stories as well as correspondence files and book reviews on close to 100 disks. Cheuse has made numerous deposits to his collection over the past two decades, with an increasing amount of born-digital content in recent years. As an interim solution, the Albert and Shirley Small Library in past years printed and interfiled the content of these electronic files while maintaining the original disks for the day when the creator's original electronic records could be likewise preserved and accessible.

As part of the AIMS project, the digital archivist was able to process collection disks using the Forensic Toolkit software and to create an EAD finding aid for the entire collection by combining multiple existing MARC records and EAD finding aids at the accession level. The individual files from the disks will be accessible through the *Hypatia* repository.

2. John Warner Papers

The vast political papers of former Senator John Warner of Virginia consist of his career as United States Senator from Virginia and Administrator to the Bicentennial from 1972-2009. Warner's collection offers an interesting insight into the composition of contemporary political collections and the intersection of born-digital assets and digitized content. Beginning in 2002, Warner's staff systematically scanned and discarded all paper-based constituent correspondence, conveying 54 CDs of what would have been hundreds of linear feet of correspondence records to the Albert and Shirley Small Library. Includes CDs containing the Senator's website.

Digital material in the Warner constituent correspondence cannot be made publicly accessible due to significant intellectual property and privacy issues. There is simply no way to obtain permissions from the hundreds of authors represented in these files. However, the disks were imaged and the content will be stored on the university's secure storage network, reducing potential preservation risks. Also, a finding aid was created for the entire collection, making both paper and digital more accessible than they were previously.

4. Yale University

Beinecke Rare Book and Manuscript Library

The Beinecke Rare Book & Manuscript Library is Yale University's principal repository for literary papers and for early manuscripts and rare books in the fields of literature, theology, history, and the natural sciences. In addition to its general collection of rare books and manuscripts, the library houses the Yale Collection of American Literature, the Yale Collection of German Literature, the Yale Collection of Western Americana, and the Osborn Collection. The Beinecke collections afford opportunities for interdisciplinary research in such fields as medieval, Renaissance, and eighteenth-century studies, art history, photography, American studies, the history of printing, and modernism in art and literature.

Manuscripts and Archives, Yale University Library

Manuscripts and Archives collects broadly in the areas of public policy and administration; diplomacy and international affairs; political and social thought and commentary; science, medicine, and the environment; legal and judicial history; the visual and performing arts; urban planning and architecture; environmental policy and affairs; psychology and psychiatry; and lesbian, gay, bisexual, transgender history and culture. In addition, the department has extensive holdings on New Haven, Connecticut, and New England history. Manuscripts and Archives also has responsibility for the Yale University Archives, the official repository for all records of the university that have enduring historical, administrative, or community significance. In addition, the department serves as the home for the Fortunoff Video Archive for Holocaust Testimonies, which currently holds more than 4,300 testimonies of willing individuals with first-hand experience of the Nazi persecutions, including those in hiding, survivors, bystanders, resisters, and liberators.

1. New Haven Oral History Project (Manuscripts and Archives)

The collection consists of digitally created audio recordings and text transcripts of oral histories conducted by the New Haven Oral History Project staff with New Haven, CT citizens. The interviews touch on a number of themes, but often focus on issues of race, class, government, education and immigration. Still growing, the collection includes more than 150 digital oral histories transferred to the archives via network transfer (no disks). Collection materials were accessioned, stored, processed, and described in an EAD record. Since this is an active collection, work will continue with the creators and through pre-custodial intervention.

2. Pelli Clarke Pelli Architects Records (Manuscripts and Archives)

A recipient of the AIA Gold Medal, Cesar Pelli and his firm have designed many of the most prominent buildings of the 20th century skyline, including the World Financial Center in New York and the Petronas Towers in Kuala Lumpur. While the complete collection exceeds 5 terabytes, initial focus will be on earlier CAD projects like the World Financial Center and the Frances Lehman Loeb Art Center at Vassar College. Just as many traditional manuscript collections that describe the evolution of a project, a book, a political career, or a scientific formula,

architectural records provide documentation of and evidence about the process of designing discrete, quantifiable objects – buildings. Born-digital architectural records provide similar insights to the design process of buildings that traditional manuscript collections provide: evidence of an initial idea, the evolution of and research into that idea, suggested modifications by editors and peers (e.g., clients), various drafts and changes as building progresses, and the publicity and marketing surrounding the final product. Preserving the various iterations – rather than just the final product – preserves an important part of our country's architectural evolution. Yale will accession, store in appropriate archival storage, and describe two architectural projects from this collection: the World Financial Center in New York City and the Frances Lehman Loeb Art Center at Vassar College. The Pelli Clarke Pelli Architects records are active collections and will continue to grow over time. Work was done during the grant period on an undescribed accession to the collection to extract metadata and prepare it for storage and access. Staff also received additional accessions of material from the firm and held two in-person records creator surveys.

3. James Tobin Papers (Manuscripts and Archives)

Correspondence, subject files, and writings documenting the professional career of the Nobel laureate and long time economics professor at Yale. A highly regarded Keynesian economist, Tobin served in both the Kennedy and Clinton Council's of Economic Advisors. Although primarily paper, the collection includes 25 3.5" computer disks. Yale will accession, store in appropriate archival storage, and describe this collection. Processing will be fully completed for the Tobin papers. The disks were imaged and technical metadata was extracted. References to the disks were added to the EAD record and were uploaded into the Hypatia application.

4. Henry Ashby Turner Jr. Papers (Beinecke Rare Book and Manuscript Library)

A long time professor at Yale, Turner is a noted historian and scholar of modern Europe, particularly Germany. The collection includes various professional writings and correspondence, including historical research data in digital form, compiled as part of a project which Turner directed to document the dealings of General Motors with Nazi Germany as GM attempted to seek evidence to counter class action lawsuits filed on behalf of victims of forced labor. The project resulted in a collection of documents (Yale's General Motors documents relating to World War II corporate activities in Europe) and a book (General Motors and the Nazis). The born-digital research data includes documentation of foreign workers at the Adam Opel AG plant in Russelsheim, Germany during the 1930s in the form of two databases (Microsoft Access and Filemaker Pro). During the grant period, the Turner papers were processed and the EAD guide was updated to include a reference to this database and both were uploaded to the Hypatia application.

5. James Welch Papers (Manuscripts and Archives)

The James Welch Papers contain manuscripts, correspondence, and personal papers documenting the life and work of author James Welch. James Welch is well known for his fiction dealing with the histories and experiences of Native Americans, and the drafts of novels and other works, together with correspondence and secondary literature, make the Welch papers a valuable resource for research in literary, American, and Native American studies. The collection spans the years 1889 to 2006, with the bulk of the collection dating from the early 1960s to

2003. This collection includes drafts of writings in digital form. The Welch papers have been previously arranged and described. The EAD guide was uploaded to the Hypatia application

6. Love Makes a Family Foundation (Manuscripts and Archives)

The Love Makes a Family (LMF) records consist of email correspondence, bylaws, reports, meeting minutes, research data, publications, Web pages, social media account files, topical files, interviews and testimonies, photographs, audiovisual recordings, and newspaper clippings documenting the history, structure, and activities of LMF, Inc. and its related organizations, the LMF Political Action Committee (PAC) and the LMF Foundation. LMF's principal goals were to pass a second-parent adoption law; support efforts to pass a domestic partnership package for state employees; defeat Defense of Marriage Amendments (DOMAs) both to the state statute and the state constitution; and pass a marriage equality law for same-sex couples in Connecticut. As the first three goals were reached by 2000, the records primarily document LMF's efforts on behalf of marriage equality. This collection includes both paper and digital records that were accessioned and processed during the AIMS grant period. The digital records consist of approximately 36 gigabytes in a variety of formats, including email correspondence, topical files, audiovisual material, photographs, websites, and social media content. An EAD guide was created and was uploaded to the Hypatia application.

Appendix E: Sample Processing Plans

I. University of Hull: Stephen Gallagher Processing Plan



Processing Plan

Acc No: 2010/15 Ref: U DGA

born-digital archives

OVERVIEW	
Collection Title:	Stephen Gallagher
Creator / Depositor:	Stephen Gallagher
Related Material at HUA:	
<p>Paper archives already deposited</p> <ul style="list-style-type: none"> - 2008/10 (42 boxes) – mainly paper with a few boxes of publications, copies of DVDs etc - 2010/14 (12 boxes) – further publications (foreign editions etc) and production material <p>Not tackled – blog / website (possibly recommend the British Library Web Archive) and email</p>	
Brief Description of the material:	
<p>Material relates to his writing, (short-stories, novels, radio and screen) including research process, drafts etc. Also material relating to his blog / website with some publicity/promotional material. There are only isolated email messages (no mailboxes).</p>	
Extent:	13.6 GB
No of files:	14,320 *
Comments re extent:	
<p>There are also 39 3" Amstrad discs</p>	
ARCHIVAL DESCRIPTION	

Proposed level of archival description to be applied:		
<ul style="list-style-type: none"> Primarily at series level 		
Justification:		
<p>Stephen Gallagher considers each piece of work as a discrete project. Interest in the material is likely to be on two accounts:</p> <ul style="list-style-type: none"> writing process following a particular story from idea through research, drafts, pitching and completion (whether publication of novel or filming of screenplay etc) a particular piece of work <p>This means that if describe the project we do not necessarily need to describe particular content</p>		
Cataloguing Priority for this accession:		Priority Score:
1. Research potential	3	18 / 24
2. HHC specialist area	3	
3. Topicality / time crucial	1	
4. UoH teaching potential	2	
5. Education potential	2	
6. Community/outreach potential	1	
7. Summary list is sufficient	3	
8. Complexity of cataloguing	3	
Scoring: 3 = high, 2 = medium, 1 = low 0 = no potential		
APPRAISAL		
Is appraisal necessary?		Yes No N/A
Potential for appraisal?		
Initial investigations identified very little material that could or should be appraised		
ARRANGEMENT		
Integrate with existing arrangement?		Yes No N/A
Does the current arrangement include b-d material?		Yes No N/A
Justification:		
There is considerable overlap between paper and born-digital material		
Potential arrangement issues?		
<ul style="list-style-type: none"> Paper files being catalogued at file level – need to consider implications for discovery & access To not try to describe each born-digital item but include an overview of born-digital material within the series description 		

<p>Any restricted / sensitive content?</p> <ul style="list-style-type: none"> • Some personal material (e.g., references for 3rd parties) that should be closed • Suggest that most recent work (i.e., last x years) should be closed [discuss this with SG] • <i>ResearchDocs</i> folder (1226 files in 87 folders, 14.5MB) material is mostly saved web-pages – need to consider arrangement /access issues • <i>MyRadio</i> folder (44 files, 1.85GB) recorded broadcasts can be included in the archives but are subject to copyright so should not be made available online via repository
<p>PRESERVATION</p>
<p>Media issues:</p> <ul style="list-style-type: none"> • Main body of material was selected by SG and transferred via external hard drive • There are 39 3" Amstrad discs that cannot be read with current hardware
<p>Content issues:</p> <ul style="list-style-type: none"> • 291 files in <i>FinalDraft</i> format (*.fdr) contact Mary-Jane Dickenson (Drama) to use their copy of <i>FinalDraft</i> – looked at files (June/July 2011) and created PDF copies for access • How to present the old website content to users as web pages (via a web browser etc) rather as individual unlinked pages
<p>Proposed preservation actions:</p> <p>Import the FinalDraft PDFs and attach to the original *.fdr file</p>
<p>Plan produced by: Simon Wilson Date: 13th Sept 2011</p>
<p>Suggested Review Date:</p>

2. Stanford University: Gould Processing Plan

Stephen Jay Gould papers.

Bio/Scope & Content:

Influential American paleontologist, evolutionary biologist and historian of science, Gould began his career at Harvard University in 1967 and worked until his death in 2002. One of the most popular science writers of our time, he is the author of 22 books, 479 peer-reviewed scholarly papers, 300 essays, and 101 reviews.

Scenario in 2009:

At the time of the AIMS grant, the Gould collection consisted of 8 accessions acquired between 2004 and 2010. Totalling over 500 linear feet of material, the collection contains specimens and legacy computer media. Items (159) of computer media were "recorded" during the accessioning process. Since then, we have uncovered more computer media (21 more sets of computer punch cards) in the 2008 accession and odds and ends scattered within folders throughout the accessions.

Media enumerated initially consisted of: 60 5.25-inch floppy diskettes, 81 3.5-inch floppy diskettes, two cartons of computer punch cards and 3 computer tapes from 1987, 1988, and 1994. The diskettes contain bibliographic databases and working drafts of many of Gould's publications. The punch cards and the data tapes appear to contain datasets used in his evolutionary biology research.

There are no online guides to any of the collection although rough container lists were created when the collection was packed up initially. The papers, audio & video are being processed concurrently.

Catalog record states: "Collection in process but open for research. Some materials may not be available. Preliminary container list available."

Trials/Actions taken:

Capture:

8 sets of punch cards (from one carton) were migrated by Computer History Museum, Mountain View, California and stored on DVD. This DVD was labeled Computer Media #144. One small set of punch cards was unreadable because there was no sorting key. Three computer tapes and 6 cartons of punch cards have not been migrated at this time (approx. 24 sets). Diskettes were labeled and numbered beginning with "Computer Media 001" or cm01. Photographic images of the diskettes and existing labels were taken for subsequent access by users.

First trial:

Disk images of floppy diskettes were created using ImageTool and a Catweasel in FRED. [A Catweasel is just an interface card for computer which don't have a floppy interface in the motherboard. Write-blocking is enabled by putting a tape at the "write-protect" area in a 5.25 inch floppy disk.] However, ImageTool did not generate an audit log file to confirm successful imaging nor a file listing of the disk contents. Our second attempt utilized an old personal computer with on-board floppy disk controller was used to image the diskettes using free software called FTK Imager. Outputs from FTK Imager include: disk images, audit log files to confirm successful imaging and file listings of the diskette contents. [Peter could not find a motherboard with floppy disk controller and interface on sale in May 2010 when I tried to do the imaging. So he brought his old computer in to do the imaging. He discovered the Gigabyte motherboard which had a floppy disk interface in Feb 2011 and built the capture station that winter.]

These were stored in a stand-alone personal computer. After detecting and cleaning computer viruses using Sophos Anti-Virus, the files were transferred to Stanford Powervault (a secured server with regular backup schedule). Only “cm94” (a high-density, 3.5-inch diskette) contained a virus which was removed.

Unreadable media (loss was 6%): CM001-CM003 (single-sided single-density 5.25-inch diskette) unreadable with existing equipment; no files copied. CM035 (double-sided high-density 5.25-inch diskette) sustained physical damage before transfer to Stanford and no files copied.

Processing

First Trial - Processing using Windows Explorer:

Quickview Plus was used to view the content of the files. Folders were created that mirrored “series” and “sub-series” in EAD and files were moved from original media folder into appropriate place using Windows Explorer. This however changed metadata associated with the files – such as original file path, etc. Adobe Acrobat Professional was used to convert files in obsolete file formats such as WordPerfect, MS DOS Word, etc. to PDF/A for access. The PDF/A version of the original files provide files with current format which can be accessed with current software. This version of the original files do not contain the original file creation dates. The file creation dates of the PDF/A files are the dates when the files were converted. The conversion also alter the last accessed dates of the original files.

Second Trial - Processing using AccessData FTK:

Logical images were created the second time around. After hearing from the curator that Creeley had deleted files on purpose that he did not want kept, Peter created logical images of the files on the floppy diskettes.

FTK extracted technical metadata (file size, creation, last modification and last accessed dates, file format, checksum, etc.) of the files in the disk images loaded. “File Category” provided a summary of how many files are in different file formats. The interface to hide the duplicate files was activated so that users are working on unique files (FTK uses the checksums of the files to identify duplicate files.). Restricted content such as credit cards, social security number, student grades, etc. were identified using the pattern & full-text searches functions. The files identified were flagged as “Privileged”. Although the search may not find ALL “Restricted” contents, it is a much better alternative to read all files. Bookmarks were created with names mirrored “series” and “subseries” in EAD. The embedded viewer (reads over 200 file formats) was used to view files with obsolete file formats. Files are then assigned to bookmarks according to intellectual contents individually or in batch. Although FTK did not forbidden the assignment of one file to more than one bookmarks, the system would change the color of the file name and its associated metadata from black to purple after the file was assigned to one bookmark. This could act as a reminder that which files had been assigned to bookmarks. “Labels” were used to represent access restrictions, document types, computer media type, and subject headings. Reports in XML/HTML format are generated to export files to access repository (Hypatia). The files carried the bookmarks, labels, privileged flag, and technical metadata with them.

EAD draft excerpts (see below)

Outstanding AIMS Project work:

- Data modeling for Gould data and metadata including EAD
- Complete EAD description for b-d materials (currently listed as Series VI – Scope & Content, Arrangement and Physical Description notes only)

- Determine delivery of b-d material, possibly by file format? – files and vehicle (Hypatia)
 - Text: manuscript writings, correspondence
 - Data sets: will be described as part of the Gould finding aid in a separate series [?] and include a live link to their digital surrogates, which will be deliverable as individual file downloads.
- Determine use/delivery of photographic images of original media labels if any
- Publish online guide in September along with paper components
- Awaiting capture of last batch of punch cards from CHM
- FOUND: 5 more cartons of punch cards as of June 2011 in 2008 accession – need to codify methodology for reading punch cards – either 1) work out exchange with CHM & quicker turn around, 2) use CHM equipment to read ourselves, or 3) costs for outsourcing
- A selection of born-digital materials will be delivered via Hypatia (demo instance)
- The catalog record will be updated with links to online guides and born-digital instance

Non-AIMS Updates

- Gould's papers will be fully processed by 8/31/11 – including all artifacts and specimens.
- The online guide to the papers will be posted online at Stanford and the Online Archive of California with series level description re born-digital materials and link to:

Series VI: Stephen Jay Gould Born-Digital Material

PHYSICAL DESCRIPTION: 52 megabytes (1,180 files)

FILE TYPES AND FORMATS

File Types: Computer Program; Data set; Document; Spreadsheet. File Formats: ASCII Text; WordPerfect 4.2, 5.0, 5.1, 6.0, 6.1; Microsoft Word 2.0, 6.0, 97, 2000; Microsoft RTF; Microsoft Excel 4.0; Lotus 1-2-3 2.0

FINDING AID LINK: To cite or bookmark this finding aid, use the following address:

<http://hdl.handle.net/10079/fa/>

Access

Collection is open for research; digital material is available online; other materials must be requested at least 48 hours in advance of intended use.

File types and formats

File Types: Computer Program; Data set; Document; Spreadsheet. File Formats: ASCII Text; WordPerfect 4.2, 5.0, 5.1, 6.0, 6.1; Microsoft Word 2.0, 6.0, 97, 2000; Microsoft RTF; Microsoft Excel 4.0; Lotus 1-2-3 2.0

Scope and Contents

This series consists primarily of the born digital material from the Stephen Jay Gould (SJG) papers. The born digital material was stored in floppy diskettes, tapes and punch cards. The original labels, if any, on the computer media are

in many cases too brief to identify the contents of the diskettes. The processor viewed the contents of each file to determine to what category the file belonged. Since SJG divided his works into "Articles", "Abstracts, Reviews, Letters, etc.", "Natural History Column", and "Books" in his bibliography, the processor followed this arrangement and added "Bibliography & Curriculum Vitae", "Teaching", "Rare Books", "Punch Cards", "Misc.", and "Computer Media Photos" as other subseries.

Details of the ten categories of files are as follows (these are added as LABELS in FTK and will display as FACETS in Hypatia):

- Articles (99 files)
- Abstracts, Reviews, Letters, etc. (107 files)
- Natural History Columns (171 files)
- Books (drafts of 12 books written by SJG in 404 files):
 - The Structure of Evolutionary Theory,
 - Full House,
 - The Book of Life,
 - Triumph and Tragedy in Mudville,
 - Dinosaur in a Haystack,
 - The Burgess Shale and the Nature of History,
 - Time's Arrow, time's Cycle,
 - The Lying Stones of Marrakech,
 - Eight Little Piggies,
 - Hidden Histories of Science,
 - The Hedgehog, the Fox, and the Magister's Pox
 - The Mismeasure of Man
- Bibliography & Curriculum Vitae (44 files)
- Teaching (12 files)
- Rare Books (28 files)

- Data Sets (11 files)
Re: computer programs and data migrated from one box of punch cards. Data in another box of punch cards is not migrated. [21 more sets discovered in 2008 addenda; unread]
- Miscellaneous (18 files) - divided into 3 sub-groups:
 - National Science Foundation (NSF)
 - Paleontological Society
 - Miscellaneous
- Computer Media Photos (165 files)

Processing Information:

Logical images of the files in floppy diskettes were created using FTK Imager and stored in a standalone personal computer. After detecting and cleaning computer virus using Sophos Anti-Virus, the cleaned files were transferred to Stanford Powervault (a secured server with regular backup schedule).

FTK Toolkit was used to assign access rights, identify restricted materials, assign series subseries information and other descriptive metadata, and generate technical metadata (MD5 checksum, file format, etc.) The files with all the metadata have been transferred to Hypatia (Hydra Platform for Access To Information in Archives).

All files will be ingested into the Stanford Digital Repository (SDR; a dark digital archive) for long term preservation. One box of punch cards was migrated by Computer History Museum, Mountain View, California, USA and stored in DVD. The DVD is assigned as Computer Media #144. One small set of punch cards was unreadable because the sorting order of the cards were mixed up. Three computer tapes and one other box of punch cards have not been migrated at this time.

Unreadable media: Computer Media #1-3 (Single sided single density 5.25 inch. floppy) unreadable with existing equipment; no files copied. Computer Media #35 (Double sided high density 5.25 inch. floppy) physical damage; no files copied. Computer Media #39 (Double sided double density 5.25 inch. floppy) blank diskettes. Computer Media #60, 134, 135 (High density, 3.5 inch. floppy) blank diskettes. Computer Media #94 (High density, 3.5 inch. floppy) contained virus and was cleaned using Sophos Anti-Virus.

3. University of Virginia: Cheuse Papers Processing Plan

University of Virginia
 Processing Plan
 Collection 10726, The Papers of Alan Cheuse

Collection Name:	The Papers of Alan Cheuse
Collection Date:	Ca. 1950 – 2009
Collection Number:	10726; accessions _ through al
Extent (pre-processing):	83 disks (3.5" and CD) approx. 5.31 MB; ca. 80 linear feet
Types of materials:	3.5" disks and CDs, video cassettes and DVDs, paper manuscripts
Custodial History:	Alan Cheuse placed the papers on loan to the Library beginning in 1987. Earlier accessions were then purchased in 2003 with a commitment to purchase further groups.
Restrictions from Donors:	Explicit digital rights have yet been discussed. Four series (Accessions 17, 18, 20, and 21) are restricted from access until 2012.
Separated Materials:	Disks have been separated from the manuscript drafts and are stored with the other media and a/v.
Related Materials:	None
Preservation Concerns:	None
Languages other than English:	None
Overview of Contents:	This collection consists of the papers of the American author, book reviewer, and George Mason University professor, Alan Cheuse. These papers include manuscripts for articles, speeches, interviews, and short stories; book reviews; screen plays; cassette tape recordings; computer disks; video cassette & DVD; printed material; contracts and royalties; passports; photographs and drawings; correspondence; research material; short stories by other authors; appointment calendars; short stories and book manuscripts.
Existing Order and description:	<p>Sixteen of the thirty-two accessions have been processed separately, as per institutional practices. They are described in both EAD finding aids and MARC records. They are each organized by type of writing (correspondence, topical files, novel manuscripts, review manuscripts, etc.) to the folder level.</p> <p>The other 16 accessions are recorded in MARC records at varying degrees of detail, some with no more than a title, date, and generic note. All computer media has been separated, numbered, and is referenced in finding aids and records, but has mostly not been processed. The contents of some disks were printed and filed with paper manuscripts.</p> <p>Seven of the accessions contain computer disk materials. Only one of these accessions has been described in an EAD finding aid.</p>

<p>Desired Processing:</p>	<p>All computer media should be processed. Additionally, all accessions should be combined into a single finding aid. Where EAD exists, these records will be combined into a single <archdesc> and <dsc> with each accession being represented as a series. The accessions represented by MARC records will be converted to series components. In addition, subject headings, which were not included in the original EAD, should be added from all MARC records.</p> <p>No further work will be done with paper materials at this time.</p> <p>The processor will create disk images of the disks and then process using FTK. Disks containing commercial works that were used for research purposes should not be imaged or stored at this time. Individual files will be labeled with the disk number so that they may later be associated with the correct container element in the EAD. Titles of individual works will be added to the finding aid so that some reference to the works available on the disks is present. This is to match the level of processing of the paper manuscripts, which are indicated by name within the collection descriptions.</p> <p>Files containing confidential information will be completely restricted at this time. Obsolete file formats will not be migrated at this time, but this work should be considered in the future. Access to materials on the disk will be at the individual file level. After imaging the disk a copy of the image will be transferred to the StoreNext preservation store. Copies of the unrestricted files will be added to the Hypatia repository for public access.</p> <p>The disk images will be referenced by identifier number within the ead. They will exist as individual subcomponents of the accession or sub-series (if it exists) and the disk number will be referenced in a "unitid" attribute. The finalized finding aid will also be uploaded into the Hypatia repository and the individual files will be linked to the accession or container they belong to.</p>
<p>Next steps</p>	<p>Reprocessing all accessions into one collection arranged intellectually, rather than intellectually within individual accessions, is recommended for the future when the collection is deemed "complete." As technology and infrastructure develop, migration of obsolete formats and redaction within restricted files in order to make them available should also be undertaken.</p>
<p>Notes to Processors:</p>	<p>Examine the contents of the CDs later in the series to determine which are simply copies of commercially produced works and do not need to be imaged.</p>
<p>Anticipated Time for Processing:</p>	<p>5 days</p>

4. Yale University: Tobin Collection Processing Plan

Processing Work Plan

Institution: MSSA

Archivist: Mark A. Matienzo

Date: June 7, 2011

Collection title: James Tobin papers

Creator: Tobin, James

Current call number(s): MS 1746, **Accession 2004-M-088**

Provenance: Gift of Elizabeth Tobin, 2004.

Extent: 8.75 linear feet; 27 3.5" inch diskettes (35.7 MB)

Overview:

Research strengths: correspondence regarding professional activities; working and final drafts of conference papers, periodical columns, and other publications.

Types of electronic records present: Correspondence (e-mail and computer-written letters); writings; spreadsheets and graphs; office files (biographical statements, calendars, publication lists, etc.), course materials. Files are primarily WordPerfect and Lotus 1-2-3; some Quicken files exist; e-mail is in text form, either in Eudora mailboxes individually saved text files.

Significant preservation concerns: See file formats above. Most significant concern is Lotus 1-2-3 files; several should be considered compound objects with graphs and formatting information.

Description:

Current: Minimal. Labels from individual diskettes have been transcribed as component titles within finding aid.

Proposed enhancement: Description should follow executed organization as specified below.

Recommended description work for later: see under organization.

Organization:

Current: Hard to determine. Paper records do not seem to have a coherent overall organization, with the exception of the correspondence; however, correspondence is still scattered between "Letters to Jim," "Professional Correspondence," "Nobel Prize Correspondence," and "Personal Correspondence." Writings are very disorganized;

Diskettes appear to be used as transfer media for files between his office, his home, and his cottage in Wisconsin. A few disks, or sets thereof, show some grouping based on type of records, such as "office files" (publication lists, telephone lists/address books) and letters that Tobin wrote in WordPerfect. Writings are not grouped together thematically.

Proposed arrangement: Arrangement should be based on record types. Within the electronic records for this accession, logical groupings and subgroupings are as follows:

- Correspondence, 1992-2001 and undated
 - Correspondence written using WordPerfect, 1992-2000
 - E-mail, 1996-2001 and undated
- Course materials for Economics 480B, 1998
 - Lotus 1-2-3 spreadsheets, 1992-1997

- “Primer” spreadsheets and graphs, 1996-1997
- Office files, 1995-2001
 - Biographical statements
 - Calendars
 - Lists of Tobin’s publications
 - Quicken files
 - Recommendation letters and lists of recommendations
 - Telephone lists
- Writings, 1992-2001

Of all groupings, the Writings grouping would need the most considerable organization and description. In the short term I recommend either not listing individual files, or listing individual files with filename and date only.

Recommended arrangement work for later: Combine paper records and electronic records into a common arrangement. Considerable attention to Tobin’s personal papers is needed, especially those related to his military service. Arrange writings alphabetically by title, identify explicit drafts, and reconcile against publication lists included in this accession as available from the Cowles Foundation. In the long term, we should plan to process the collection as a whole and integrate all the accessions into a common arrangement.

Appraisal:

Diskettes 1-3, 11, and 17 should be discarded; #1-3 contain printer drivers; #11 contains modem software; and #17 contains many deleted files and is mostly blank.

Some of Tobin’s “office files” are of uncertain or low research value, such as the Quicken files, biographical statements and telephone lists. The publication lists are of questionable value as the Cowles Foundation has a detailed publication list in PDF form; however, Tobin has some topic-specific publication lists that may be helpful. Some of the office files also appear to be inventories of paper files, which may or may not be reflected in the paper records previously acquired.

Restrictions:

Other (paper) correspondence within this accession is restricted. E-mail contains both personal and professional correspondence; personal/family correspondence includes reference to health issues. Consider restricting e-mail under similar conditions. Most letters written using WordPerfect are professional in nature. Recommendation letters and Quicken files (which deal with Tobin’s personal finances) should be restricted.

Preservation:

Proposed action now: Investigate migration options for Lotus 1-2-3 files, particularly those that reference graphs.

Recommended for later: Migrate WordPerfect files to PDF/A; migrate e-mail to a different format.

Access:

See Preservation. Files should be extracted into a storage option such as the YUL Rescue Repository so they can be paged on request. This collections does not have a high level use, so there is probably not an immediate need to create use copies.

Appendix F: Policies, Templates, Documents, etc.

I. AIMS Donor Survey

[Institution Name] Digital Material / File Survey – Part I

Revision: May 6, 2011 (revised AIMS July 7, 2010 version)

Note: This part of the survey is designed to be a prompt sheet for phone / face-to-face interview with donors by curators / digital archivists.

1. General Work & Computing Habits

- 1.1. What are your chief activities? (e.g. writing, research, lecturing, other professional activities)
- 1.2. What kinds of records do you create, maintain, and use in the course of each of these activities? (e.g. drafts of writings, research notes, lecture notes, journals, diaries, correspondence, photographs, databases, etc.)
- 1.3. Can you describe your general work habits with computers in support of these activities? (e.g. you write first by hand, then input work into computer; you use different computers for different kinds of work, you're always online, etc.).

2. Digital Material Creation

- 2.1. Are you solely responsible for creating your digital files?
- 2.2. If not, who else is involved, and what are their roles?
- 2.3. Do you maintain digital files created by others? If yes, how do you separate your files and files created by others?
- 2.4. Do you share your computer with other people? If yes, how are files created by different people separated?
- 2.5. Do you separate your personal files from your work files?
- 2.6. What are the earliest and latest creation dates (roughly) of your digital files?

3. Varieties of Digital Material

- 3.1. What types of digital files are created? (e.g. word processing files, images, spreadsheets, databases, etc.)
- 3.2. If you create files in both digital and paper formats, do certain files exist in both formats? (e.g. drafts of writings, email, etc.)

4. Digital Material Organization

- 4.1. How are digital files named?
- 4.2. Is some kind of version control used? (e.g. filename1, filename2, to represent 1st and 2nd drafts of the file.)
- 4.3. How are your digital files currently organized? (e.g. filed in named folders? by projects? by topics? some other scheme?)
- 4.4. Have you always had this organization? If not, can you summarize/characterize previous organizations, and roughly when and why you made changes?
- 4.5. Are digital files destroyed in regular intervals?
- 4.6. Do you use more than one computer (e.g. office desktop, office portable computer, home desktop, etc.)? If yes, how do you synchronize files between different computers?

5. Mobile Device

- 5.1. Do you use smart phones (e.g. Blackberries, iPhone, Android phone, etc.)? If yes, do you store contents in the smart phone elsewhere?
- 5.2. Do you use tablets PC (e.g. iPad, etc.)? If yes, do you store contents in the tablet PC elsewhere?

6. Email

- 6.1. Do you have multiple email accounts?
- 6.2. Which email programs/services are you using? (e.g. Email program provided by your work place, Outlook, Mac Mail, Hotmail, Gmail, Yahoo! Mail, etc.)
- 6.3. How is your email currently organized? (e.g. in self-created email folders, etc.)
- 6.4. Have you always had this organization? Do you use the sorting function with any regularity to re-order your email?
- 6.5. How is email saved? (e.g. untouched in the email program, a copy in your PC, printed out in paper, etc.)
- 6.6. Are email and paper correspondence managed together or separately?
- 6.7. Do you use address books?
- 6.8. Is there a space quota assigned to your email account? If yes, have you ever exceeded the quota assigned?

7. Calendar Software

- 7.1. Do you use calendar software with your computer(Outlook, Google Calendar, 30 Boxes etc.)? Which one?

- 7.2. Do you use calendar software in your mobile device?
- 7.3. Do you have any synchronization issue between the calendars in your mobile device and your computer?

8. Webpages / Blogs

- 8.1. Do you have webpages / blogs?
- 8.2. Are webpages / blogs updated? How often?
- 8.3. What software do you use to update webpages / blogs?
- 8.4. Have copies (digital or paper) of previous versions been kept?

9. Social Networking Sites (e.g. Facebook, LinkedIn, Twitter, etc.)

- 9.1. Do you have social networking accounts?
- 9.2. Are account information (e.g. profiles, photos, etc.) updated? How often?
- 9.3. Have copies (digital or paper) of previous versions been kept?

10. Photo / Video Sharing Sites (e.g. Flickr, Picasa, YouTube, etc.)

- 10.1. Do you post photos / videos to these web sites? If yes, which one?
- 10.2. How often do you post contents?
- 10.3. Do you delete photos / videos posted? If yes, do you have a copy of the deleted postings?

11. Document Sharing Sites (e.g. SlideShare, Scribd, Google Doc, etc.)

- 11.1. Do you post documents to these web sites? If yes, which one?
- 11.2. How often do you post contents?
- 11.3. Do you delete documents posted? If yes, do you have a copy of the deleted postings?

12. Digital Files Storage / Backup

- 12.1. Do you / your institution have a backup routine for your files / emails? If you don't know, do you mind we ask your technical support? How can we contact your technical support?
- 12.2. What media are used for backup files? (e.g. optical disk, hard disk, file server, web based backup service such as SugarSync, etc.)
- 12.3. Do you transfer files in your old computer to your new computer? If yes, what types of files are transferred? Did you encounter any problems in transferring the files?
- 12.4. Do you keep your old computers? Roughly when were they being used? Can you tell us what platforms they run on?
- 12.5. Have you ever experienced a serious hardware failure (e.g. hard-drive crash)? If yes, are the files in the affected computer recovered?

I2.6. Are any digital files stored in unusual storage media? (e.g. punch cards, 8 inch. floppy diskettes, etc.)

I3. Privacy and security

I3.1. Are some digital file types of a sensitive nature? (e.g. tax records, medical records, peer-review comments, letters of recommendation, student records, etc.)

I3.2. Are there files that you would want destroyed? If yes, please provide details so that we can act upon when we encounter such files when processing your files.

I3.3. Do any digital files require passwords?

I3.4. Where are user names and passwords kept? What service / software are used to save them?

I3.5. Do you use digital watermarks? On what types of digital files? For what reasons?

I4. File Transfer Arrangement

I4.1. Do you want to delete any files / re-organize the files before the transfer?

I4.2. Are there files you would like to transfer to us later? When?

Institution Name Digital Material / File Survey – Part II

Revision: June 21, 2010

Note: This part of the survey is designed to be filled out by digital archivists regarding technical details of the tools used to create digital material.

1. Hardware

- 1.1. List the hardware configurations of each computers / mobile device. (e.g. manufacturer, model no, cpu, ram, hard drive capacity, video card, etc.)
- 1.2. Find out if the computers have USB ports or CD writers which could be used to copy the digital files.

2. Software

- 2.1. List the operating system and other system software with version no., installed in all the hardware (in 1).
- 2.2. Check if system date and time are set correctly. List the “Time Zone” used, if any.
- 2.3. With the help of the donor; list the main application software, with version no., used to create digital files.
- 2.4. If Microsoft Office is used, find out if the “User Name” field is set to the name of the donor. Find out similar setting for other main application software used.

3. Internet Access

- 3.1. Find out if the digital archivist can use the Internet access in the donor’s office using the digital archivist’s portable computer?

4. Networking

- 4.1. With the help of the donor; confirm if the computer is connected to file servers. Confirm if the donor save files in the file server. How much file server space is used by the donor?

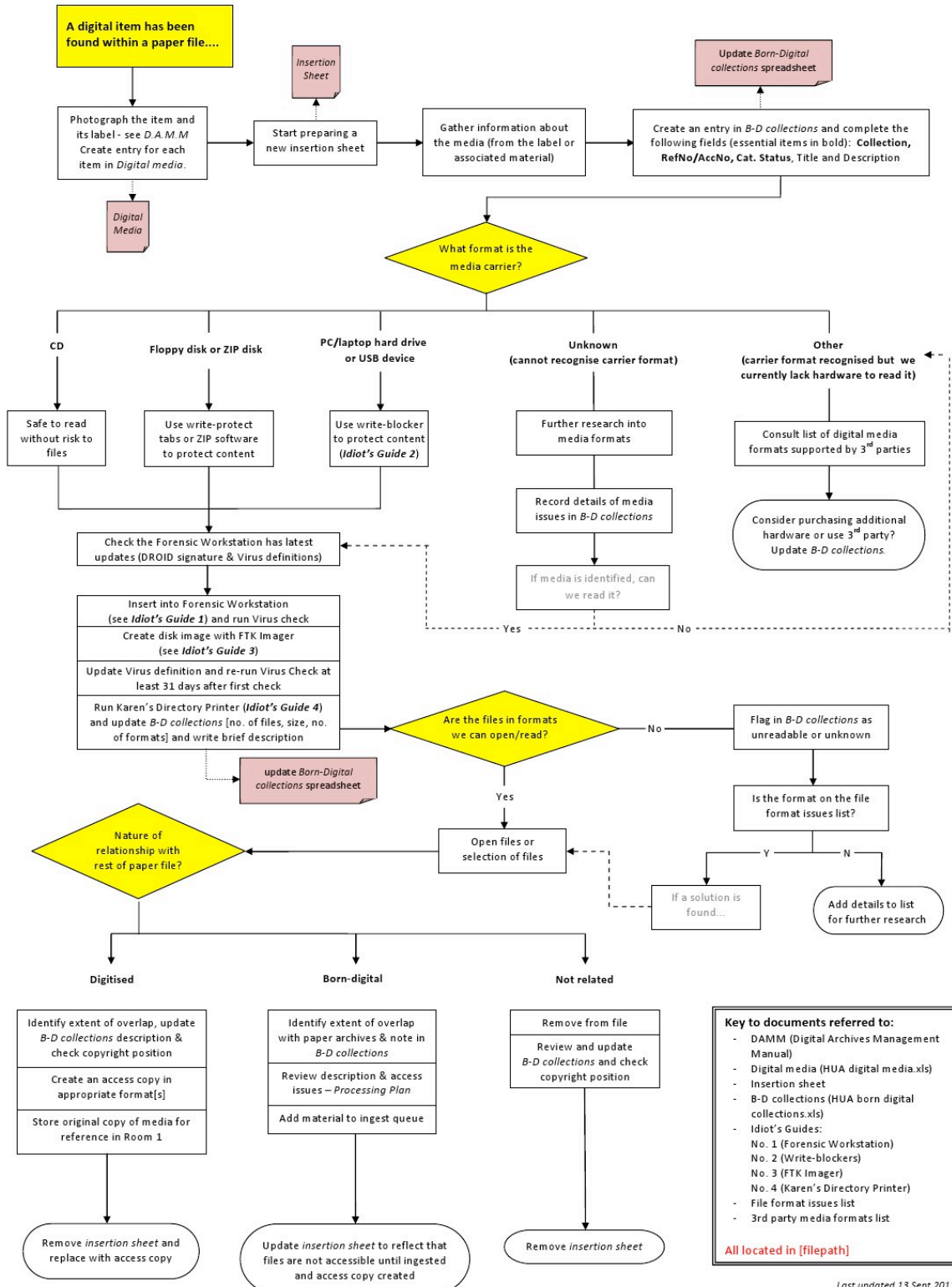
5. Security

- 5.1. With the help of the donor; confirm if login is required to access desktop computers / mobile devices?
- 5.2. With the help of the donor; confirm if a digital certificate is used by the donor to login / sign digital files / encrypt digital files?
- 5.3. With the help of the donor; confirm if digital files are encrypted?

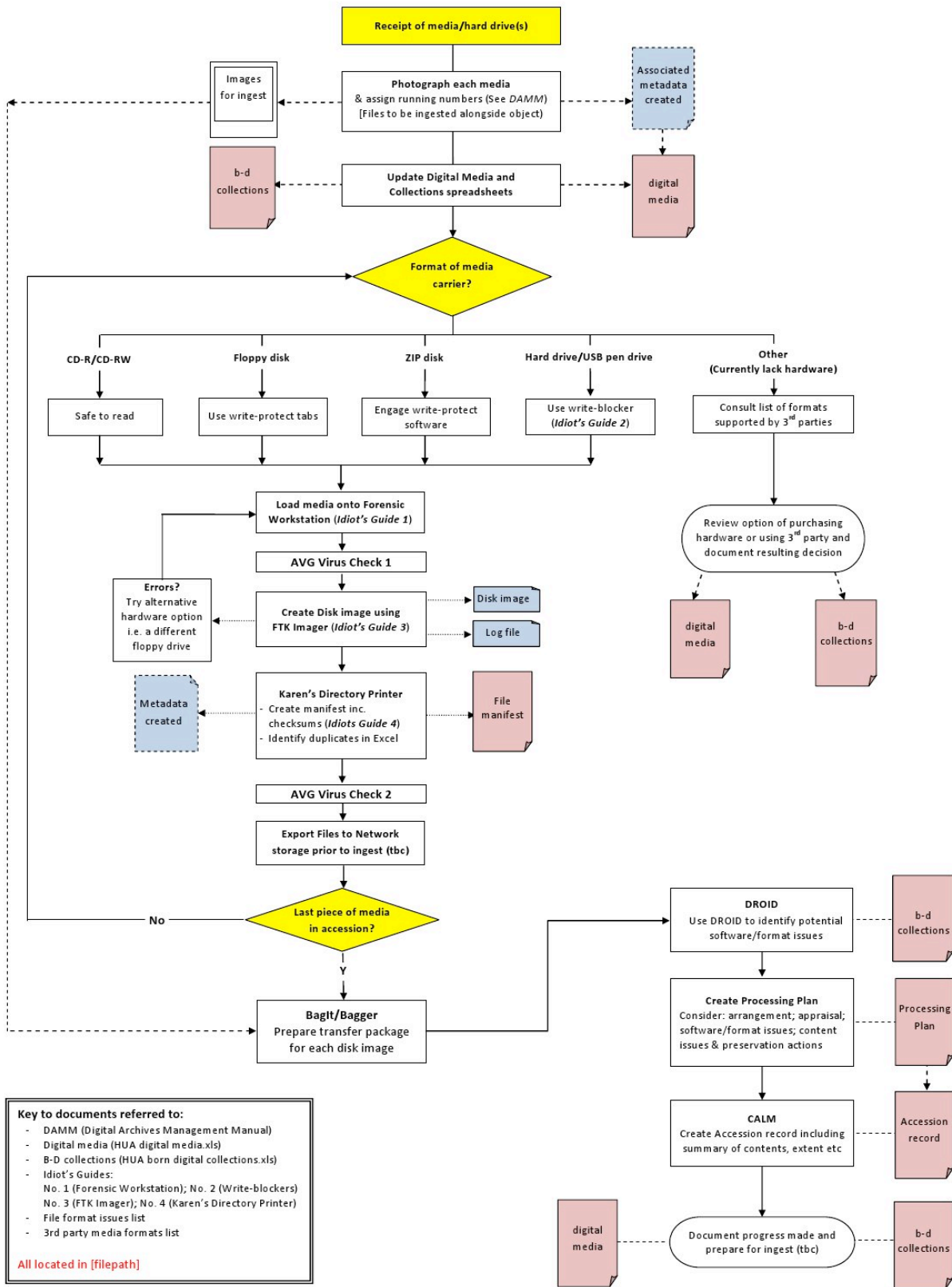
2. University of Hull Accessioning Workflows



Workflow 1: Discovery of digital media within paper files



Workflow 2: Accessioning Born-Digital Archives





3. University of Hull Digital Media Photography Form

Acc/ Ref No		Item No	
Aspect:	Front / Reverse / Side / Case /		



4. University of Hull Insertion Sheet

This Insertion Sheet has replaced a digital media item which has been removed from this file. The content is not currently accessible because:

1. The media on which it is stored is obsolete and we currently lack the hardware to read it
and/or
2. The file formats contained on the media are obsolete or very rare, possibly requiring specialist software to read them.

As part of our continuing work with born-digital archives we are working on accessing the content and migrating the files to newer formats. Our intention is to provide access to the content in an appropriate form, but this is an ongoing process.

OVERVIEW	
Description from catalogue: File. Police, Lay Visiting Information Sheet. Including 5" floppy disk, "Lay Visitors" draft info sheet on "Word Perfect",	
DETAILS	
Brief description of the material, including quantity and media formats (if known): One 5.25" floppy disk	
Content (if known): Content inferred from label, which reads: 'Draft Visitors Draft Info sheet. Rec'd from Richard Cal-land, 25/1/89'.	
Media issues? Obsolete media; we currently lack the hardware to read it. We are considering purchase of hardware or 3rd party services.	File format issues? Content possibly created in WordPerfect; may need migrating.
Notes: The disk may have been read when first received as the cataloguer has noted that the format is WordPerfect	
Insertion sheet completed by: Nicola Herbert	Date item removed from file: 28 May 2011

5. Guidelines for Creating Agreements at Stanford University

When a repository decides to begin active collecting of born-digital materials, it should review its current agreements to ensure that issues specific to the acquisition, preservation, and delivery of born-digital content are fully addressed and are consistent with overall institutional policies and requirements. Direct consultation with the receiving institution's legal counsel, is strongly advised. Two general things to remember are 1) it should be based on your institution's policies whether they are stated or implicit and 2) there are some things that are better recorded as an 'attachment' or addendum to the agreement.

It is also important to remember; these changes to standard legal agreement templates are not retroactive; when you deal with legacy data, you will need to determine your course of action based on a review of the original agreement and your ability to revisit the issue with the copyright holders. To ensure that legal agreements arrived at remain current within the evolving environment of intellectual property law and institutional policy and practice, it is prudent to periodically review your legal agreements with curators, administration, archivists and legal counsel.

Examples below are drawn from the current deed of gift template at Stanford University Libraries' Special Collections. Examples below are drawn from the current deed of gift template at Stanford University Libraries' Special Collections.

1. Ownership:

Repositories still receive much digital content transferred by physical media, however, there is a growing trend to receive virtual transfers — these might include a drop box, institutional network, self-deposit, etc. Therefore, all references regarding the material being transferred should refer to the “donor” as the owner of both physical and digital materials.

e.g. [Donor Name] (“Donor”), the owner of the physical property [and digital materials] described below [and as added to from time to time], hereby gives, transfers and conveys to Stanford University (“Stanford”) all the donor's title and interest to the following materials to become part of Stanford University Libraries.

2. Exclusivity:

Your repository may want to consider a statement about exclusivity even though it may be difficult to enforce. One issue that might arise: a repository might be offered files from a dealer and will need to have a reliable method to determine if they have already received them. To ensure that you hold the “originals” and the rights to have the only copy — granted from either the creator or his/her heirs/assigns — would be important to document.

e.g. The Collection will be placed exclusively with Stanford.

3. Transfer or ownership and materials (method/date):

Special note might be made regarding transfer method or time for born-digital materials, which may not come at the same time, or in the same manner; associated paper files. At Stanford, arrangements about physical trans-

fer of materials often are documented in appendices and correspondence, rather than stated in the agreement itself.

e.g. Ownership of the Collection will vest with Stanford; and title to any Collection placed after the date of this Agreement will transfer to the Stanford on delivery. Digital material will be transferred by [METHOD*] on [DATE].

4. Preservation:

Mention should be made regarding your repository's plans for digital preservation in addition to storage and preservation for analog materials. This is not a promise to deliver the digital files in perpetuity.

e.g. Stanford will exercise the same degree of care over the preservation of the Collection as over the preservation of similar property which is kept in the Stanford University Libraries. (e.g. climate-controlled storage for physical materials and digital preservation in the Stanford Digital Repository for digital materials.)

5. Duplicate, Redundant or Out of Scope Materials:

While this clause does not currently address digital files, it is something that should be covered in conversations with the donor:

e.g. Stanford reserves the right to return, or, with the consent of the Donor, to discard/destroy any duplicate or redundant material or any material not deemed of archival value.

6. Restrictions:

Restricted materials are usually referred to in an agreement but detailed in accompanying documentation.

e.g. To guard against violation of confidentiality or the use of the Collection to harass, injure, or damage, Donor reserves the right to restrict access to specific portions of the Collection ("Private Material").
[Choose one: Such material has been identified on Attachment [A] OR Donor agrees to identify such material for Stanford before the materials are physically transferred].

7. Metadata and discovery (finding guides, etc.)

There should be some statement(s) granting permission to describe the materials — both analog and digital. This statement would also cover copyright to this new description.

e.g. The Donor explicitly permits Stanford to create finding guides to the Collection and full-text search for unrestricted digital material as well as associated metadata required for the preservation and description of the Collection. Stanford will own the copyright in any technical or descriptive metadata added during the course of processing. The Donor shall be provided with a copy of any such finding guides upon request.

8. Delivery agreements

As stated previously in this white paper, a repository should not promise either to preserve or deliver all born-digital content — especially if it is taking in new formats or those not "currently" supported by your digital preservation repository or the preservation community.

e.g. Stanford will provide access to the Collection pursuant to its policies and procedures, which are online at [website]. Unless provided otherwise in this Agreement, Stanford is under no obligation to provide access to the Collection. In no event, is Stanford obligated to provide access to all or part of the Collection if doing so would cause financial (such as costly restoration) or health and safety concerns (such as documents with mold). Additionally, Stanford's providing access to the Collection must be done in compliance with copyright laws.

9. Lastly, although not exclusive to born-digital content, the agreement should cover permission to post digital material via the web. This inclusion would only cover materials (either born-digital or digitized) in which the donor held copyright. Your institution should have policies in place regarding issues of discoverability, access and use, i.e. having the agreement to post does not imply that researchers be given the ability to download without registering with the site or seeking permission and approval.

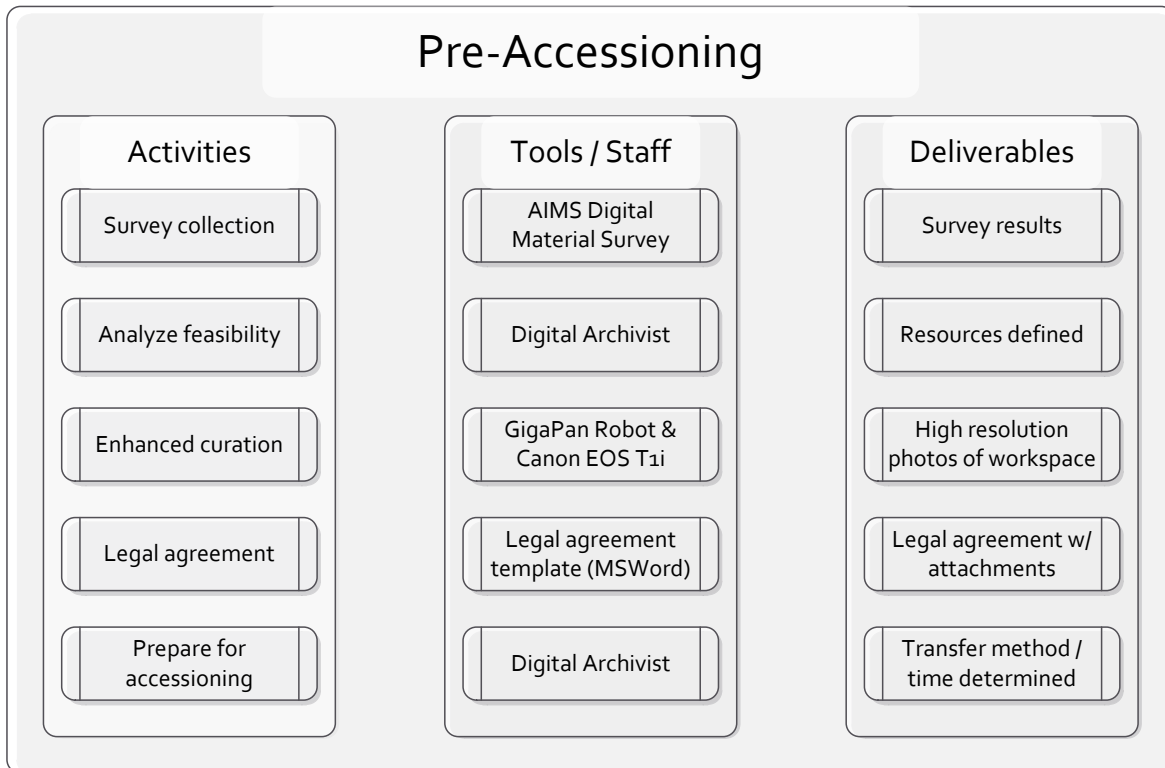
e.g. [Choose one option from the below. Option A should be used when Donor is the copyright holder for the Collection, and is assigning that copyright to Stanford. Option B should be used where the Donor is retaining copyright, but granting Stanford a license for use. It is anticipated that Option B will be most common. Option C should be used where Stanford receives the materials only, and has no rights to reuse. Option C should always be used where Donor is a collector and has no copyright interest in the materials.]

OPTION A: Donor hereby assigns, as part of this gift, all of the intellectual property rights, including but not limited to copyrights that Donor may possess in the Collection. Donor understands that he is forever and irrevocably granting to Stanford all exploitation rights in the Collection, including but not limited to the sole and exclusive right to publish all unpublished writings and copyright the same in all media now known or hereafter created.

OPTION B: No rights to any copyright in the Collection are being transferred to Stanford. Donor hereby grants to the Stanford an irrevocable perpetual royalty-free [exclusive] license to use and exploit the works of the Collection for which the Donor has copyright, individually or collectively for educational and not-for-profit purposes. This [exclusive] license includes the right to copy the works of the Collection or published materials for which the Donor holds copyright, collectively or individually, for educational and/or not-for-profit purposes in all media now known or hereafter created, including but not limited to print, audio, electronic, video, optical disc, photographic, digital and film. Without limiting the foregoing, to the extent not prohibited by copyright, the Stanford University Libraries is permitted to post a digital copy of the works of the Collection either collectively or individually, on Stanford University websites.

OPTION C: No rights to any copyright in the Collection are being transferred to Stanford.

6. Stanford University Processing Workflow



Introduction:

The born-digital materials workflow at Stanford University Libraries’ Department of Special Collections & University Archives has undergone many changes over the course of the AIMS project. In this set of four workflow documents, we have aligned current tools and deliverables into a matrix focused on archival and curatorial activities. It is not meant to be exhaustive nor to cover all file formats – this workflow is primarily for textual formats, such as text files, spreadsheets, databases, etc. (For example, a separate workflow for stewarding digital photography collections is currently under development. With many of our workflows there will be significant overlap.)

Notes for Pre-Accessioning:

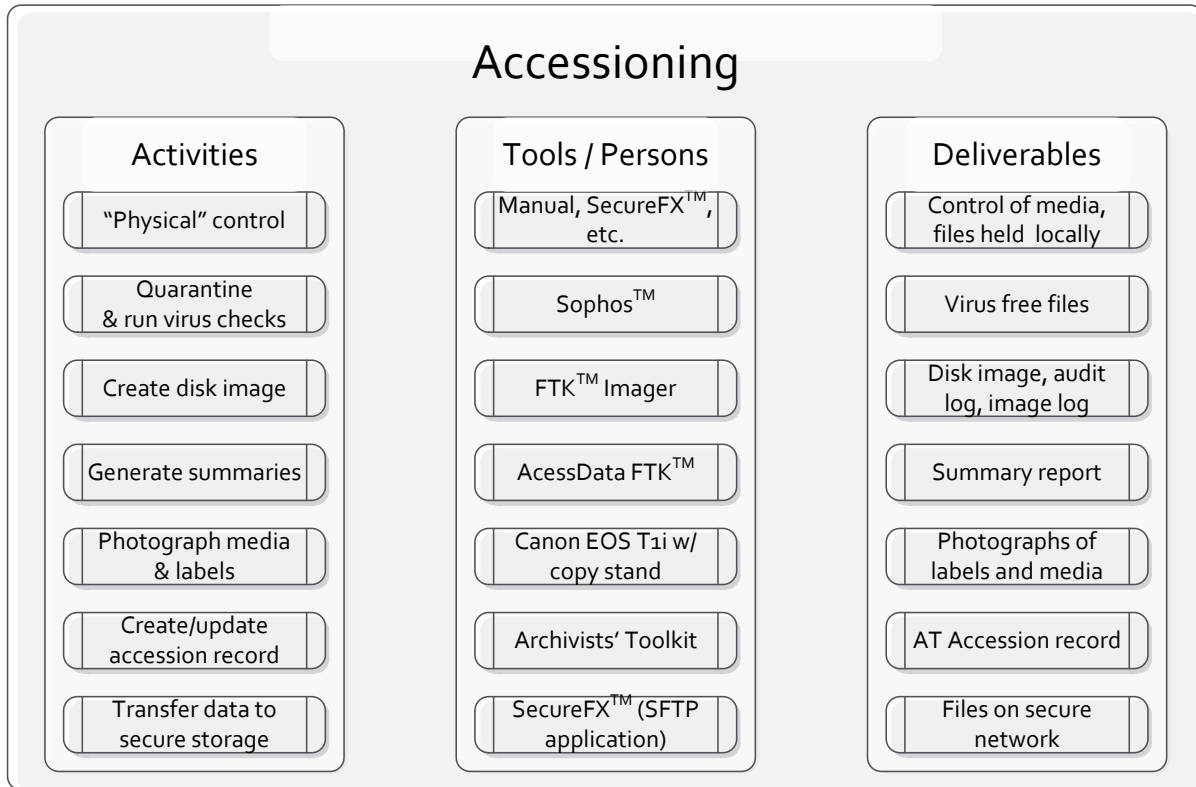
An archivist – likely the digital archivist should assist with analyzing the feasibility of taking in this material. This might encompass performing a test capture, previewing material in situ with donor/curator, researching hardware, software, cloud services; obtaining price quotes for capturing from unusual formats, and speculating on rates for capture and processing time. All of this is documented and reported to head of the archival unit and the cognizant curator. It will become part of a package often used in support of acquiring a collection.

This feasibility study and documentation will inform archival and curatorial staff as to the costs associated with both capturing and processing the digital materials. In addition it will provide the digital archivist and technical staff with necessary information before transfer regarding the need to allocate more network space, purchase special equipment, etc.

Enhanced curation might take on many forms, such as documentation of the creator’s work habits (oral history), recollection of documents or documentation of their workspace. Above we have noted our capacity to create high-resolution photographs of a creator’s workspace using a camera and robot. [www.gigapansystems.com]

The legal agreement is negotiated between the curator and the donor. The curator might seek advice or input from archival staff, particularly the digital archivist. The template for the legal agreement should be reviewed every few years and should acknowledge born-digital materials and their stewardship. Since the legal agreement will often cover multiple accessions, applicable documentation regarding a specific accession can be appended to the agreement or later accessions.

SULAIR Born-Digital Workflow, FY12 – Peter Chan, Glynn Edwards



Notes for Accessioning:

Physical control entails either counting, labeling, and numbering actual physical media or gaining control via virtual transfer, such as copying files from a network or picking up from a drop box. At this time, physical media should be entered into a media log to track success rates and loss during capture. This log allows us to track loss statistics for various types of media and by collection.

Our current policy is similar to that described in Beinecke’s policy guidelines. We create a forensic (bit-for-bit) disk image which is then transferred with accompanying technical and descriptive metadata into the Stanford Digital Repository for preservation storage. In general, these disk images are not delivered but will remain in storage until the files / collection is processed. Capturing a forensic disk image allows us to keep associated files in case they are needed during processing or for enabling delivery. One such example is of proprietary fonts that might be used by designers. If they are not captured, it is impossible to accurately render these files in any virtual environment down the road.

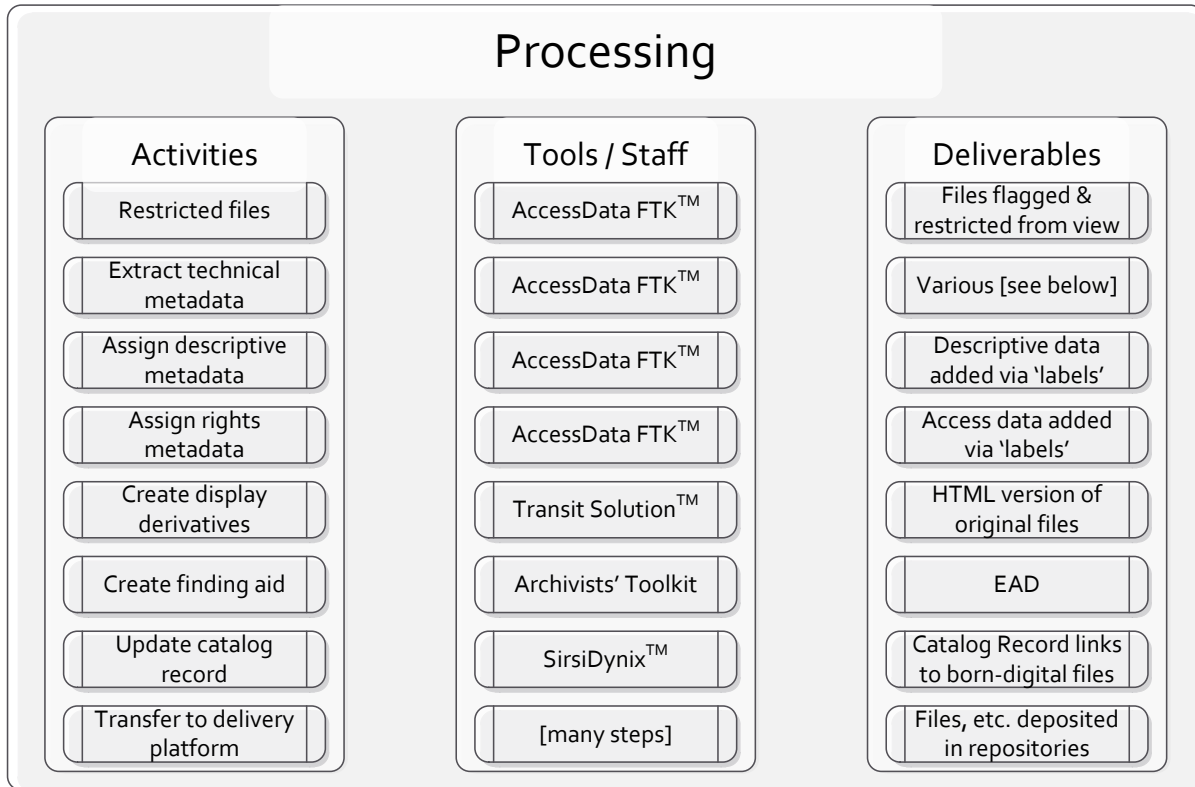
Our current tools are a mix of open-source and commercial software. Some tools, like FTK Imager, are commercially

developed but free to download and use. Special Collections currently uses Archivists’ Toolkit (AT) for its collections management and for EAD creation. We are using Imager to capture files (which is free) and AccessData’s FTK software for processing (which is not).

Similarly, we use a mix of equipment to perform captures from legacy and modern media. We have a Forensic Recovery Evidence Device – FRED – and home-built PCs for capturing floppy and zip disks [see also <http://born-digital-archives.blogspot.com/>]

In addition, we have made the decision to photograph media. While they do serve as an image of the artifact, they are primarily a record of the metadata written onto the media label. While often it may be cryptic or irrelevant (if media has been overwritten), it is similar to a folder title and may be the only external hint of the creator’s organization. It also alleviates the need to have original media available during processing. These photographic images are packaged together with the disk image and technical summaries. The files and metadata are exported from FTK with a script that creates xml files for ingestion into our Fedora repositories (preservation and delivery).

SULAIR Born-Digital Workflow, FY12 – Peter Chan, Glynn Edwards



Notes for Processing:

For collections that are primarily textual rather than image or sound/moving image, our main tool for arrangement and description is AccessData's FTK (Forensic Toolkit). Since FTK is designed for law enforcement, it has some capabilities that are not useful for archival processing, but it offers aspects that are very useful. FTK allows the processor to:

- See the original organization/structure of the collection, whether the data results from a capture of hard drives or files from physical media
- See the technical metadata associated with the files
- Assign metadata to the files – either individual files or in bulk

(see <http://www.youtube.com/watch?v=hDAhbR8dyp8>) Associated metadata might fall into many categories. Descriptive metadata might include keywords or subjects and intellectual arrangement. This is done with the use of the "label" function in FTK. We are working on standardizing our labels so they map more easily into our repositories at the end of this process.

Extract technical metadata

Examples of technical metadata extracted are checksums (noting duplicates); file formats; file creation, last accessed and modification dates.

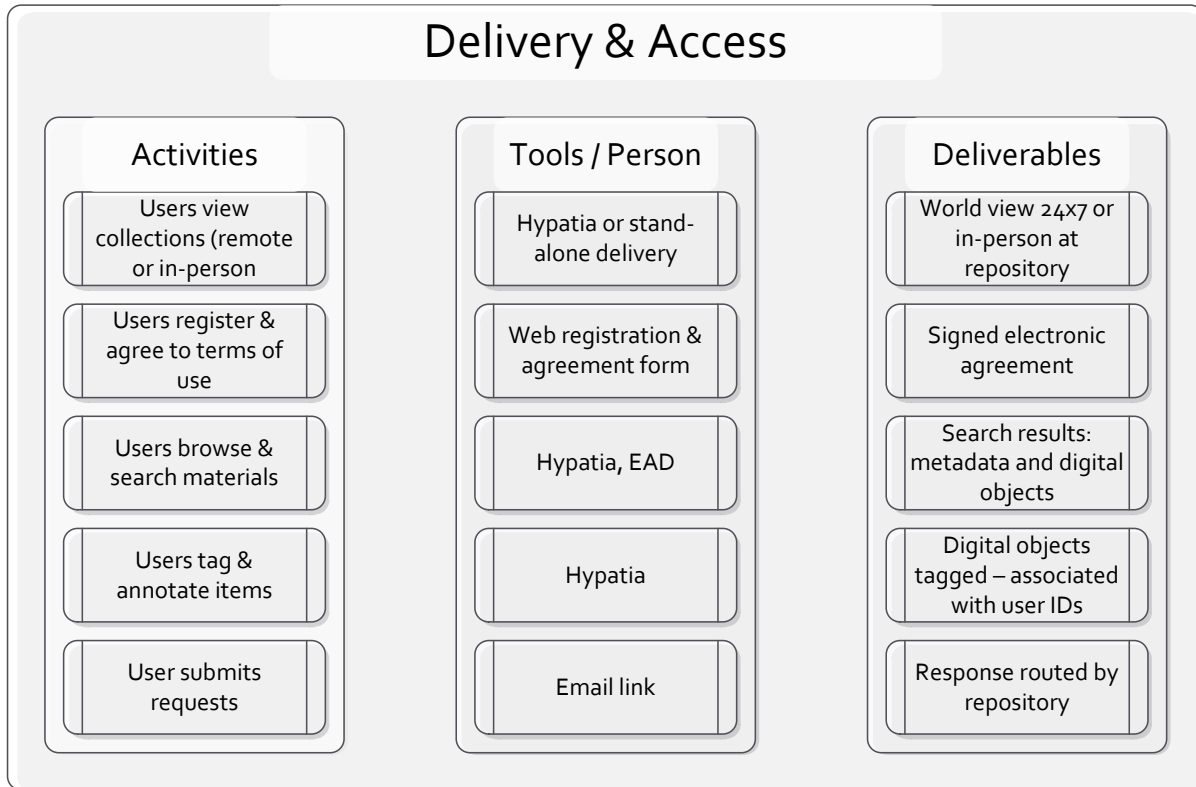
Create finding aid

Once the files themselves are processed in FTK, the archivist will create basic EAD description using the Resources module in Archivists' Toolkit. We have made a conscious choice to not describe born-digital materials at the file level but rather at the series level – i.e. it will not appear in-line with other descriptive metadata in the finding aid. This is for two reasons. First, the amount of born-digital material makes this approach unsustainable – this is especially true of legacy material which would need to be integrated into old guides. Second, current and future development points to the merger of metadata – from analog materials (EAD) and digital materials (FTK) into one search platform. The ability to search across metadata and full-text would obviate the alternative need to spend resources merging that data by hand in an EAD-only environment.

Transfer to delivery platform

This box covers many steps needed to extract data from FTK and AT and ingest it into the two digital repositories. An XSLT transformation program – developed by Peter Chan and Bess Sadler – exports the data and files from FTK to Fedora; SecureFX is used during the transfers; and current work on scripting EAD into Fedora and ingesting the packages into Fedora are underway.

SULAIR Born-Digital Workflow, FY12 – Peter Chan, Glynn Edwards



Notes for Delivery & Access:

How users discover digital materials and how they access it are two important issues that are important to consider throughout the process of stewarding born-digital content.

Hypatia

Software for delivery and access, developed as part of the AIMS project, is called Hypatia and it will use components of the Hydra technology stack being developed as part of the Hydrangea project. The files and metadata will be full-text searchable.

Outcomes of this effort at the end of September will be:

- four Stanford AIMS born digital collections will be processed (data/metadata from FTK & EAD will be delivered)
- processed AIMS collections will be in a Stanford hosted Hydra Fedora environment
- creation of a Hypatia Hydra head that supports the deposit, accessioning and delivery of these collections

For more information about Hypatia, see Appendix H.

During the processing stage, the archivist will determine where and how the digital objects will be delivered and to whom. The Discovery and Access function (2.4) details

many of the options, such as via the web or in-person via reading room computer, open to the world, the Stanford-community, or to a specific IP address or individual – such as the donor and archivist.

Discoverability will be from multiple avenues and the depth of that discovery will depend on where the user request originates. Since the catalog record – uploaded to OCLC – will describe and link to the born-digital content, a user will be redirected to the collection in Hypatia. If the user search originates in the EAD finding guide posted online at Stanford University Libraries and the Online Archive of CA, there is fuller description – especially of accessioning and processing notes – and, again, a link to the content and all associated metadata, both created by archivists and inherent in the files, that take the user to Hypatia.

One outcome for Hypatia, perhaps not developed by the end of the grant period, is to allow patron driven tagging and annotation both to the digital files and the descriptive metadata. Users will be able to download certain files in a variety of ways, again set by the archivist. For special requests, they will email their request and it will be reviewed and acted on by the digital archivist.

SULAIR Born-Digital Workflow, FY12 – Peter Chan, Glynn Edwards

7. Beinecke Library Born Digital Archival Acquisition Collection and Accession Guidelines

BEINECKE RARE BOOK & MANUSCRIPT LIBRARY BORN DIGITAL ARCHIVAL ACQUISITION COLLECTION

- DRAFT -

The Beinecke Library (BRBL) is committed to collecting, preserving, and providing access to important literary archives including materials documenting creative processes, writing lives, aesthetic communities, publication records, etc. in a range of formats and media. In keeping with this commitment, the Library recognizes and appreciates the increasing and inevitable significance of born-digital materials in literary archives. We have established, therefore, a flexible framework for working with archive creators and their representatives in various contexts to systematically, efficiently, and safely work with born digital manuscripts, correspondence, and related materials as they are acquired, accessioned, organized, maintained, accessed, and used for various research and education purposes.

To that end, the Beinecke Library employs the following guidelines in approaching the assessment, evaluation, collection, capture, accession, and preservation of materials created using digital media;

- BRBL collects digital archival materials in any and all relevant formats (including text, image, sound, etc);
- In acquiring born digital materials, a forensic approach, including 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;
- Because BRBL is interested in collecting digital materials that have substantive research value, such materials may be segregated from other materials in a broadly-conceived digital archive (spam and other commercial email, for example, may be excluded; extensive personal image or sound file collections may be curated by BRBL before collection and accession). This more limited acquisitions approach will be applied primarily in cases where a small group of materials are to be acquired (a specific body of correspondence, for instance) and not in the case of acquisition of a complete archive;
- In order to retain whatever organization, file structures, and associated data exists in the a digital archive or collection, BRBL staff members need direct access to digital files in their original environment to perform data appraisal, capture, and verification; it is suggested that representatives of archive creators (family and friends, book dealers, agents) should not manipulate, rearrange, extract, copy etc. data from its original source in anticipation of offering the materials to BRBL for gift or purchase.

Appendix G: Technical Evaluation and Use

I. AccessData FTK3.3

Purpose

AccessData FTK (Forensic ToolKit) generates summary information on a collection (single floppy disk or a collection with floppy, zip, CD and hard disks) of files and provides different views of files, sophisticated search, bookmarking and labeling functions.

Use of Software

This software can be used in the accessioning, arrangement and description phases of the AIMS framework for born digital material.

Key Functionality

1. Summary information of a collection (single floppy disk or a collection with floppy, zip, CD and hard disks) by file extension, file category, file status and Email message.
 - a. Summarizes files by their extensions, such as .TXT, .JPG, and .DOC and lists them in a tree view.
 - b. Summarizes files by type, such as a word processing document, graphic, email, executable (program file), or folder, and lists them in a tree view.
 - c. Summarizes files by status such as deleted files, duplicate items, and encrypted files, etc. and lists them in a tree view.
 - d. Provides message counts of Emails in AOL DBX, PST (Outlook email), NSF (Lotus Notes email), MBOX (Thunderbird, Netscape, Eudora, etc. email) formats.
2. Different views of files, including explorer tree, file list, file content and thumbnail.
 - a. Explorer Tree View lists directory structure of disks/folders, similar to the way one would view directory structure in Windows Explorer in original order.
 - b. File List View displays files and pertinent information about files, such as filename, file path, file type, file formats (identified by FTK) and checksums (generated by FTK), etc.
 - c. File Content View displays files as Hex (hexadecimal representation), Text (in different character encoding scheme such as ASCII, Chinese Traditional (Plane 1), EBCDIC (37 United States), Mac OS

Roman, Windows 1252 (Latin I), etc.), Filtered (file's text created during indexing), and Natural (file's contents as it would appear normally) formats. The "Natural" format uses the Oracle Stellent INSO filters for viewing hundreds of file formats without the native application being installed.

d. Thumbnail View displays graphics files in thumbnails in photo-album style.

3. Index Search, Pattern Search and Fuzzy Hashing

- a. Index search compares search terms to an index file containing discrete words or number strings found in a collection. Index search options include: "Stemming Words" that contain the same root, such as raise and raising, "Phonic Words" that sound the same, such as raise and raze, "Synonym Words" that have similar meanings, such as raise and lift, "Fuzzy Words" that have similar spellings, such as raise and raize.
- b. Pattern Search includes many predefined regular expressions for searching, including the following: U.S. Social Security Numbers, IP Addresses, U.S. Phone Numbers, Visa and MasterCard Numbers, U.K. Phone Numbers, and Computer Hardware MAC Addresses, etc. Users can also create their own pattern.
- c. Fuzzy Hashing is a tool which provides the ability to compare two distinctly different files and determine a fundamental level of similarity. Traditional cryptographic hashes (MD5, SHA-1, SHA-256, etc.) are useful to quickly identify known data, to indicate which files are identical. However, these types of hashes cannot indicate how closely two non-identical files match. Fuzzy hashing identifies similarity by a score from 0-100. A score of 100 would indicate that the files are close to identical. Alternatively a score of 0 would indicate no meaningful common sequence of data between the two files.

4. Provide Labeling and Bookmarking

- a. Labels give you a method of grouping files in a completely user defined way.
- b. A bookmark is a group of files that users want to reference. These are user-created and the list is stored for later reference, and for use in the report output. Users can create as many bookmarks as needed. The main difference labels and bookmarks is that bookmarks can be nested within other bookmarks and labels do not have such feature. This makes bookmark a good choice for representing "series" and "subseries". Install

Verdict

FTK is the only software I know to perform all the functionalities mentioned above in a totally integrated environment.

Further Information

<http://accessdata.com/products/computer-forensics/ftk>

Questions:

Contact Peter Chan, digital archivist, Stanford University Libraries, at pchan3@stanford.edu.

2. AccessData FTK Imager 3.0

Purpose of software

FTK Imager is a data preview and imaging tool.

Use of Software

The software can be used to create forensic or logical (deleted files, unallocated space not included) images of local hard drives, floppy diskettes, Zip disks, CDs, and DVDs, entire folders, or individual files from various places within the media in the accessioning phase of the AIMS framework.

Key Functionality

FTK Imager is a data preview and imaging tool created by AccessData Corp. With FTK Imager, you can:

- Create forensic images of local hard drives, floppy diskettes, Zip disks, CDs, and DVDs, entire folders, or individual files from various places within the media.
- Create logical images of the contents of folders. The image created will include only logical files. It will not include deleted files, unallocated space, etc. It does not store sector information.
- Preview files and folders on local hard drives, network drives, floppy diskettes, Zip disks, CDs, and DVDs.
- Preview the contents of forensic images stored on the local machine or on a network drive.
- Mount an image for a read-only view that leverages Windows Explorer to see the content of the image exactly as the user saw it on the original drive.
- Export files and folders from forensic images.
- See and recover files that have been deleted from the Recycle Bin, but have not yet been overwritten on the drive.
- Create hashes of files using either of the two hash functions available in FTK Imager: Message Digest 5 (MD5) and Secure Hash Algorithm (SHA-1).
- Generate hash reports for regular files and disk images (including files inside disk images) that you can later use as a benchmark to prove the integrity of your case evidence. When a full drive is imaged, a hash generated by FTK Imager can be used to verify that the image hash and the drive hash match after the image is created, and that the image has remained unchanged since acquisition.
- Encrypt data during export to an image.

Identified and Analyzed File Systems

- Microsoft: FAT 12, FAT 16, FAT 32, NTFS, exFAT

- Apple: HFS, HFS+
- Linux: Ext2FS, Ext3FS, Ext4FS
- Others: ReiserFS 3, VXFS, CDFS

Identified and Analyzed CD and DVD File Systems and Formats

Alcohol (*.mds), IsoBuster CUE, PlexTools (*.pxi), CloneCD (*.ccd), Nero (*.nrg), Roxio (*.cif), ISO, Pinnacle (*.pdi), Virtual CD (*.vc4), CD-RW, VCD, CD-ROM, DVD+MRW, DVCD, DVD-RW, DVD-VFR, DVD+RW Dual Layer, DVD-VR, BD-R SRM-POW, BD-R DL, BD-R SRM, CloneCD (*.ccd), HD DVD-R, HD DVD-RW DL, SVCD, HD DVD, HD DVD-RW, DVD-RAM, CD-ROM XA, CD-MRW, DVD+VR, DVD+R, DVD+R Dual Layer, BD-RE, DVD-VRW, BD-ROM, HD DVD-R DL, BD-R RRM, BDAV, Pinnacle (*.pdi), HD DVD-RAM, ISO, CD-R, Virtual CD (*.vc4), SACD, DVD+RW, DVD-ROM, VD-R, DVD-VM, DVD-R Dual Layer, DVD+VRW, BD-R SRM+POW

Verdict:

The ability to create logical image is extremely important when a bit-by-bit forensic image is not allowed.

Further information

FTK is a proprietary software but is free and can be downloaded at <http://www.accessdata.com/downloads.html>.

Questions:

Contact Peter Chan, digital archivist, Stanford University Libraries, at pchan3@stanford.edu.

3. Comparison of 5.25” Floppy Disk Drive Solutions

Purpose

Most modern computers do not have the hardware needed to read 5.25 floppy diskettes. This review compares 4 solutions to connect a 5.25 floppy drive to your existing computers or a new one built from the motherboard suggested. Catweasel is an expansion card to be inserted in the PCI slot of your existing PC. Both KryoFlux and FC5025 are bare circuit boards with a USB interface for connecting to the USB port of your existing PCs. Gigabyte GA-880GA-UD3H is a motherboard with a floppy disk controller which allows you to connect your 5.25 inch floppy drive.

Key features for each solution:

	Catweasel	KryoFlux	FC5025 USB 5.25" floppy controller	Gigabyte GA-880GA-UD3H
Hardware	PCI expansion card	Printed circuit board with USB interface	Printed circuit board with USB interface	Motherboard
Included Software	IMAGE (GUI) Command Line Tools	DTC (command line) GUI	Disk Image and Browse (GUI) Command Line Tools	Nil (the following is based on FTK Imager)
Operating System requirement	Windows XP; works on Linux with additional software	Windows XP, Vista (32-bit) Windows 7 (32/64-bit) Mac OS X Linux	Linux x86 2.6.24 Mac OS X PPC 10.4.11 Mac OS X Intel 10.6.4 Windows XP SP3 32-bit Windows 7 (32/64-bit)	Linux x86 Windows XP, Vista (32-bit) Windows 7 (32/64-bit)

	Catweasel	KryoFlux	FC5025 USB 5.25" floppy controller	Gigabyte GA-880GA-UD3H
Supported disk type / File system	<p>PC-formats (180K up to 1440K)</p> <p>Amiga DD and HD (also 5.25" formats)</p> <p>Atari 9, 10 and 11 sector disks</p> <p>Macintosh 720K, 800K, 1440K (DD, GCR, HD)</p> <p>Commodore 1541, 1571, 1581 (C64, C128 and 3.5" C-64 disks)</p> <p>XTRA High density with 2380KByte per disk</p> <p>Nintendo backup station 1600KB format</p> <p>Atari 800XL (all MFM formats, FM under development)</p> <p>Apple IIe disks (Apple DOS 3.3 and up)</p>	<p>KryoFlux supports dumping any floppy disk to "stream files", which contain the low level flux transition information present on a disk. It also supports output of a range of common "sector dumps" to allow you to use your dumped images right away in your favorite emulator. The currently supported disk image formats are:</p> <p>KryoFlux stream files</p> <p>CT Raw image, 84 tracks, DS, DD, 300, MFM</p> <p>FM sector image, 40/80+ tracks, SS/DS, DD/HD, 300, FM</p> <p>FM XFD, Atari 8-bit</p> <p>MFM sector image, 40/80+ tracks, SS/DS, DD/HD, 300, MFM</p> <p>MFM XFD, Atari 8-bit</p> <p>AmigaDOS sector image, 80+ tracks, DS, DD/HD, 300, MFM</p> <p>CBM DOS sector image, 35+ tracks, SS, DD, 300, GCR</p> <p>Apple DOS 3.2 sector image, 35+ tracks, SS, DD, 300, GCR</p> <p>Apple DOS 3.3+ sector image, 35+ tracks, SS, DD, 300, GCR</p> <p>DSK, DOS 3.3 interleave</p> <p>Apple DOS 400K/800K sector image, 80+ tracks, SS/DS, DD, CLV, GCR</p>	<p>Apple DOS 3.2 (13-sector)</p> <p>Apple DOS 3.3 (16-sector)</p> <p>Apple ProDOS</p> <p>Atari 810</p> <p>Calcomp Vistagraphics 4500</p> <p>Commodore 1541</p> <p>Kaypro 2 CP/M 2.2</p> <p>Kaypro 4 CP/M 2.2</p> <p>MS-DOS</p> <p>North Star MDS-A-D</p> <p>TI-99/4A</p>	<p>Microsoft: FAT 12, FAT 16, FAT 32, NTFS, exFAT</p> <p>Apple: HFS, HFS+</p> <p>Linux: Ext2FS, Ext3FS, Ext4FS</p> <p>Others: ReiserFS 3, VXFS, CDFS</p>
Disk image output format	Raw (plain, .bin, .d64, .d71, .d81, .adf, .xfd), .d64 with error information, .atr	Raw	Raw (.d64, .img, .po, .do, .dsk)	Raw (dd), SMART, E01, AFF
Directory listings of all files in the image	No	No	No (only browse)	Yes
Log file (date, time, checksums, actions, results)	No	No	Partial (only success/failure and bad sectors)	Yes (checksum of both original disk and the disk image)
Filesystem browse (appraisal)	No	No	ProDOS, MS-DOS and Kaypro disks	MS-DOS
Integrate with QuickView Plus (appraisal)	No	No	No; can import disk images into FTK Imager	Yes

	Catweasel	KryoFlux	FC5025 USB 5.25" floppy controller	Gigabyte GA-880GA-UD3H
Cost	USD120	USD3,000 (non-personal edition price) Euro94.95 (Personal Edition Advanced)	USD55.25 (additional USD48 for disk drive external enclosure and power supply)	USD120

Verdict

Building a pc based on the Gigabyte GA-880GA-UD3H motherboard is only solution mentioned above to allow you to use QuickView Plus and your antivirus software to see/scan the files in the floppy diskette without creating a disk image. The other 3 solutions require the user to create a disk image of the diskette and extract the disk image in order to see the files using QuickView Plus or to scan the files with your antivirus software in a floppy diskette. Anyway, all four solutions provide unique features and users have to match the solutions to their problems.

Further information:

FC5025: <http://www.deviceside.com/fc5025.html>

Catweasel: http://www.jschoenfeld.com/products/catweasel_e.htm

Gigabyte GA-880GA-UD3H: <http://www.gigabyte.us/products/product-page.aspx?pid=3758#ov>

Kryoflux: <http://www.kryoflux.com/>

Contact Peter Chan, Digital Archivist at Stanford University Libraries, at pchan3@stanford.edu for questions.

I would like to thank Mark Matienzo for supplying the information on FC5025 and commenting this review.

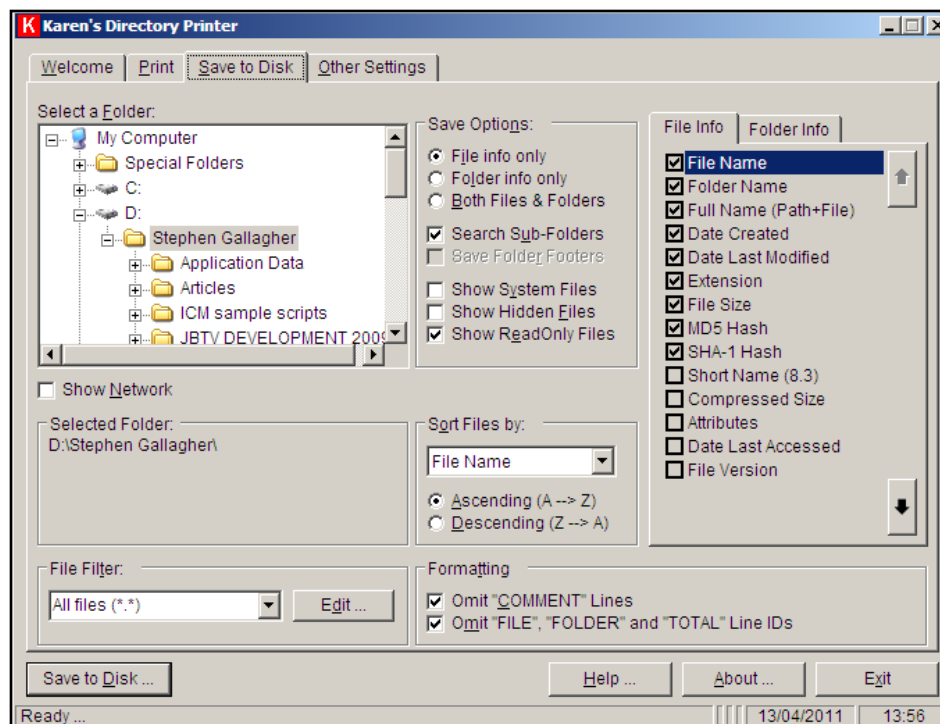
4. Karen's Directory Printer (v.5.3.2)

Purpose

Karen's Directory Printer is a freeware tool that can capture key details about each file and folder in an accession. It cannot be used to capture details from disc images.

Use of Software

We were first drawn to the potential of this software by accounts of its use by West Yorkshire Archives Service. This software can be used to create a manifest of the files that have been transferred to us before we undertake any processing, that is with-in the accession phase of the AIMS framework, and can be used in conjunction with write-blockers.



The information can include file and folder name, the full path, the size of the file (in kb), the date the file was created and last modified. For folders it can record the folder name, the number of files, the number of sub-folders and total size of the folder. This data can be saved as a text file, using .csv format, that can be easily imported into MS Excel and then manipulated in a number of ways including identifying duplicate items –where the checksums match – irrespective of the filenames.

It is possible to create file and folder information at the same time, but having two separate manifests makes data analysis and the potential for re-use easier. The software remembers the settings between uses which make subsequent re-use easier and quicker.

Key Functionality

One of the most useful features of this software is that it can create both MD5 and SHA1 checksums and these can be compared with checksums generated through other tools like FTK Imager. The ability to capture file extension also provides an indication of possible file types to be encountered - this can then be verified through the use of DROID.

The ability to view key information about the folders - and in particular the number of files and its size. This information can be used to provide a useful perspective of the entire collection and may suggest particular folders for appraisal and this can be documented in the processing plan.

Verdict

It is possible to create file and folder information at the same time, but having two separate manifests makes using the data in further tasks easier. Indeed this is one factor that has meant we have decided to keep using this software despite similar functionality being offered by FTK Imager.

Although its use involves another piece of software in our workflow we felt the tool was simple and easy to use and feel confident in suggesting its use by depositors who may wish to create a list of files that they intend to transfer.

Further Information

<http://www.karenware.com/powertools/ptdirpm.asp>

5. Curator's Workbench

Purpose

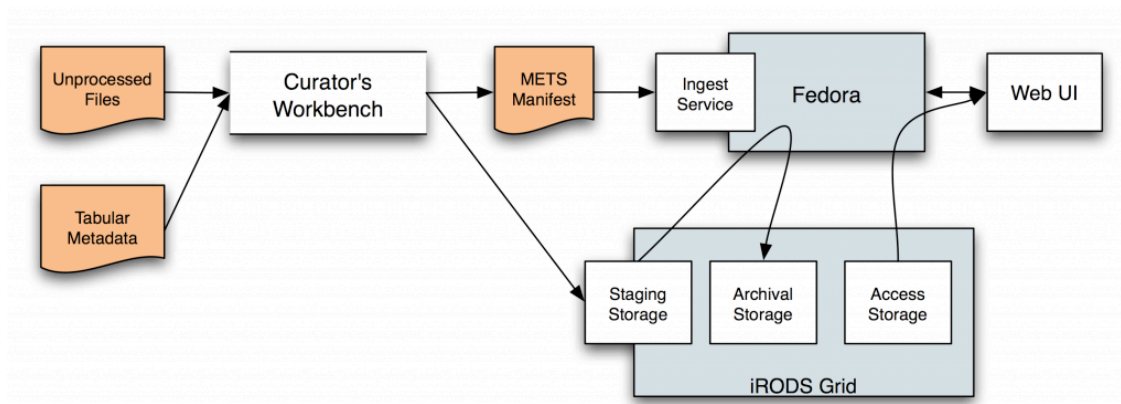
The Curators Workbench is an open source tool designed to assist with the accession, arrangement, description and staging of digital objects. The tool was developed as part of the Carolina Digital Repository over the summer of 2010 by developer Greg Jensen and Erin O'Meara Electronic Records Archivist.

The tool is still in active development with version 3.0 being due for release in early October 2011. It has been deliberately designed to have a modular framework to allow other institutions to use and extend the tool according to their particular institutional requirements. In the summer of 2011 Greg and Erin hosted a number of workshops in the UK as part of their attempts to establish a user community that can actively contribute to the development process.

Use of Software

The tool creates a METS file documenting the processes that have been applied and can create MD5 checksums and unique IDs for each object (UUIDs). MODS descriptive metadata can be mapped to individual objects and folders using the impressive crosswalk feature.

The software requires each accession to be handled as a distinct project which is useful and each project is then built around the METS manifest which tracks the objects and their metadata and is then exported to form the basis of a submission package prior to ingest.



See <http://www.lib.unc.edu/blogs/cdr/index.php/2010/12/01/announcing-the-curators-workbench/>

Key Functionality

The crosswalk is one of the distinctive features of Curators' Workbench with the crosswalk editor allowing a user to visually map their data with MODS data elements. At present this only supports tab-separated metadata sources but it is planned to be extended to any delimited file and XML sources. The ability to save and then re-use the

crosswalk definition allows a user to generate the MODS records. This re-use saves considerable time and effort and in most cases should avoid the need for custom scripts for each data source. The editor also allows you to add standard text for example a statement relating to copyright as part of the crosswalk process.

The tool also includes a staging area designed to facilitate the processing and ingest of files to your preservation storage environment, critical considering the sheer number of files contained within many born-digital archive accessions. The staging area can be configured to your specific storage environment and it can also be used to identify issues prior to forming the submission package.

The tool does claim that you can add descriptive metadata but it was unclear whether this could be applied in batch mode or even whether this conformed to any descriptive standards and it also allows you to view the properties of each file.

Verdict

The tool looks very professional and very polished and in the most part is easy to use. The crosswalk editor does require getting used to but is worth the investment in time and effort.

It is difficult to be too judgemental for a tool that is in such active development, but aspects I would like to see include / explore further are;

1. How easy it is to create suitable metadata to implement the crosswalk from a position of having a batch of born-digital files, which is how many collections will be received
2. Whether a distinction can be made between the original folders and the staged files – with both containing the same files it is easy to forget “where” you are
3. Clarification whether arranging the files is an intellectual process only, as is proposed with the Hypatia tool, before you start renaming, re-arranging and deleting objects

Further Information

Curator's Workbench at UNC that includes links to manuals, screencasts, etc.:

<http://www.lib.unc.edu/blogs/cdr/index.php/about-the-curators-workbench/>

Curators Workbench wiki:

<https://github.com/UNC-Libraries/Curators-Workbench/wiki>

Appendix H: Technical Development

I. Functional Requirements for Arrangement and Description

The functional requirements presented here were developed by the AIMS partners over several months of discussion and testing of various tools that can perform various activities within the born-digital archival workflow. The functional requirements are described in 13 overall sections. Within each there may be “Further Questions or Comments” — areas of discussion that were not decided on before the end of the grant period or that required some development work before they could be decided — and “User Stories” — examples of the proposed tool in use in a hypothetical situation. Although these requirements are unfinished and were only partially implemented in the Hypatia demo application (see *Appendix H.3*), the partners present them here so that they may fuel future work in this area.

Functional Requirements for AIMS Hydra Head (“Hypatia”)

This work is licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported License.

A&D_00: Fundamentals

The arrangement and description tool must provide a mechanism which allows an archivist to do the following:

- Define an intellectual arrangement of transferred archival records that reflects the provenance and original order of the records. The original files and directory are not moved or modified in any way.
- Create and edit descriptive metadata for those records. It must also be possible for the archivist to add descriptive data to individual files in addition to adding descriptive data for any of the given levels of arrangement.

Levels of arrangement as defined within archival practice, and accordingly, this tool set includes collection, series, subseries, folder, and item. (see **A&D_12 Overview**)

Each archival collection will have its intellectual arrangement, that is the arrangement of the material in a hierarchical nature that intends to reflect its original creation or arrangement within a recordkeeping system. Over time additional material may be received and these accessions will be integrated into the collection and the intellectual arrangement will be updated. The arrangement is used to portray and distinguish critical elements of context. Software tools like Archivists’ Toolkit and CALM allow archivists to create the intellectual arrangement with description based on content standards like DACS or ISAD(G).

Other tools might be used to create exhibitions but any organization of the material for this purpose should not be confused with the intellectual arrangement. A user is able to sort a collection into a particular order that suits them (e.g., by date) via the discovery and access tools.

AIMS partners can supply BagIt-based SIPs, either in directory or zip/tarball form. Rubymatica packages files with technical metadata from FITS/DROID into SIPs.

Further Questions or Comments

The terms used in this document are common within American practice. Arrangement terms used in the UK are collection, sub-fonds, series, sub-series, item [the unit of production; e.g., one file] and piece [pages within a volume or individual letters within a bundle etc]

A&D_01: Graphical User Interface

The arrangement and description tool must have graphical user interface (GUI) and implement and reflect best practices and conventions of user interface (UI) design. The application should operate within a web browser for best cross-platform compatibility. The tool set should be relatively easy to use and should likely reflect user interaction paradigms to which archivists are accustomed, such as those found in applications already in use by the AIMS partners (namely Archivists' Toolkit and CALM). Accordingly, in some cases these functional requirements may refer to other functional requirements, documentation, or specifications as applicable to demonstrate user interfaces requirements. Individual requirements within this document may also explicitly describe specific user interface requirements.

The original organization of the files and directories within an ingested accession and the archivist-defined intellectual arrangement have special status, and that status should be obvious in the UI and should be enforced by the UI. For example, it is essential that users authenticate as an archivist in order to modify the intellectual arrangement. (Keep in mind that a detailed description of collection permissions may not be covered by this document.)

When working on the intellectual arrangement, archivists will need ready access to technical metadata such as the original full path of a given file (see **A&D_02: Technical Metadata**). It may be useful to have a "show original" function within a contextual menu that would show the originally ingested file in the left pane.

Further Questions or Comments

It would be useful to be able to associate digital photographs of media with imported collection components. For example it would be useful, particularly for minimally processed collections, to be able to show images of the source media (floppy disks in particular) alongside the digital files it contains. These photographs must be distinguishable from actual content from the media, possibly via an explicit metadata folder or similar (this could also contain an original 'manifest' and/or web survey information if so desired).

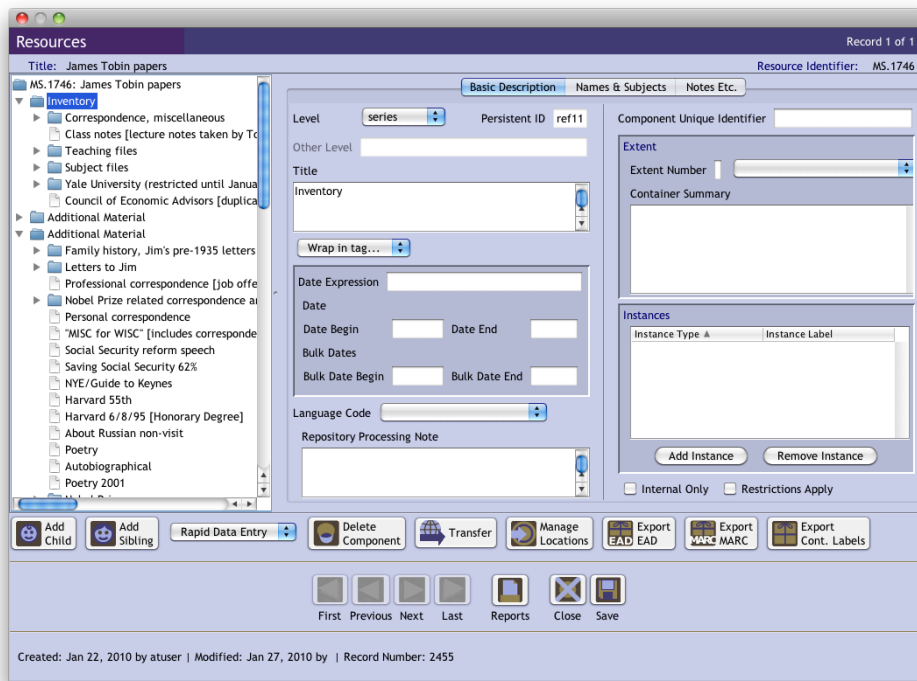
A&D_01.01: Representation and manipulation of hierarchy

The graphical user interface should allow users to view and interact with hierarchical structures representing the intellectual arrangement and the original arrangement of files and directories within ingested accessions. There should be distinct panes representing the structure of the intellectual arrangement and representing the accessions. For each component level in the intellectual arrangement, the user interface should present associated digital assets (see **A&D_01.02** and **A&D_12**) and an interface to view and edit descriptive metadata elements (see **A&D_03.02**).

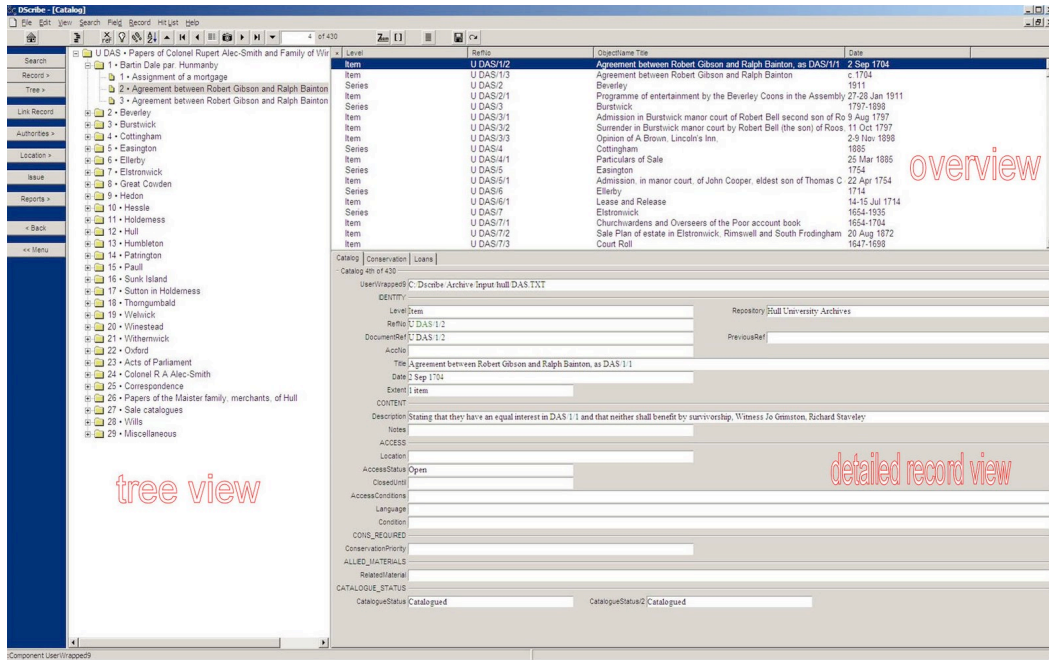
In addition, the tool should allow the following operations (applies to intellectual arrangement only unless otherwise specified):

- Collapse and expand record nodes for viewing (applies to both the original ingest and the intellectual arrangement)
- Add new child record (see **A&D_12**)
- Add new sibling record (see **A&D_12**)
- Copy all or part of an existing structure to the intellectual arrangement. Ideally, we could copy structure of the original ingest, or copy all or part of an intellectual arrangement.
- Delete a record in intellectual arrangement. This applies only to the intellectual arrangement. Recursive folder delete is a dangerous operation, and the UI must add special safe guards. We should be able to delete a record, only if it has no children in order to avoid orphan entries.

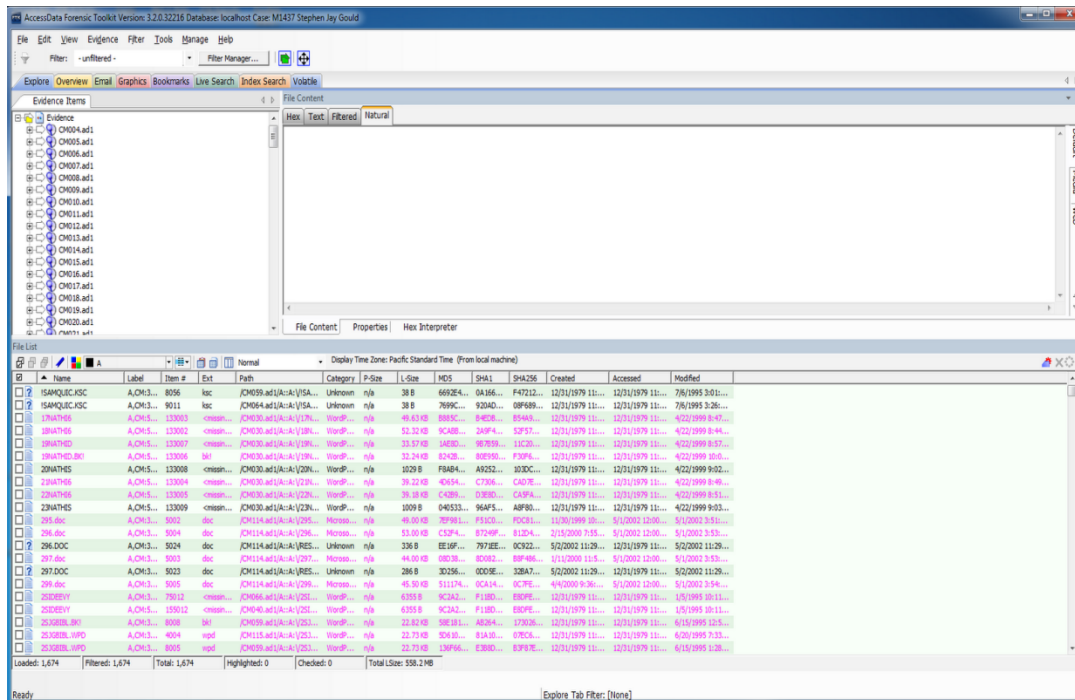
Sample Screenshots



Archivists' Toolkit



CALM



Forensic Toolkit

A&D_01.02: Drag and drop functionality

*NOTE: This is heavily interrelated with **A&D_12**. Please refer to functional requirements in detail below.*

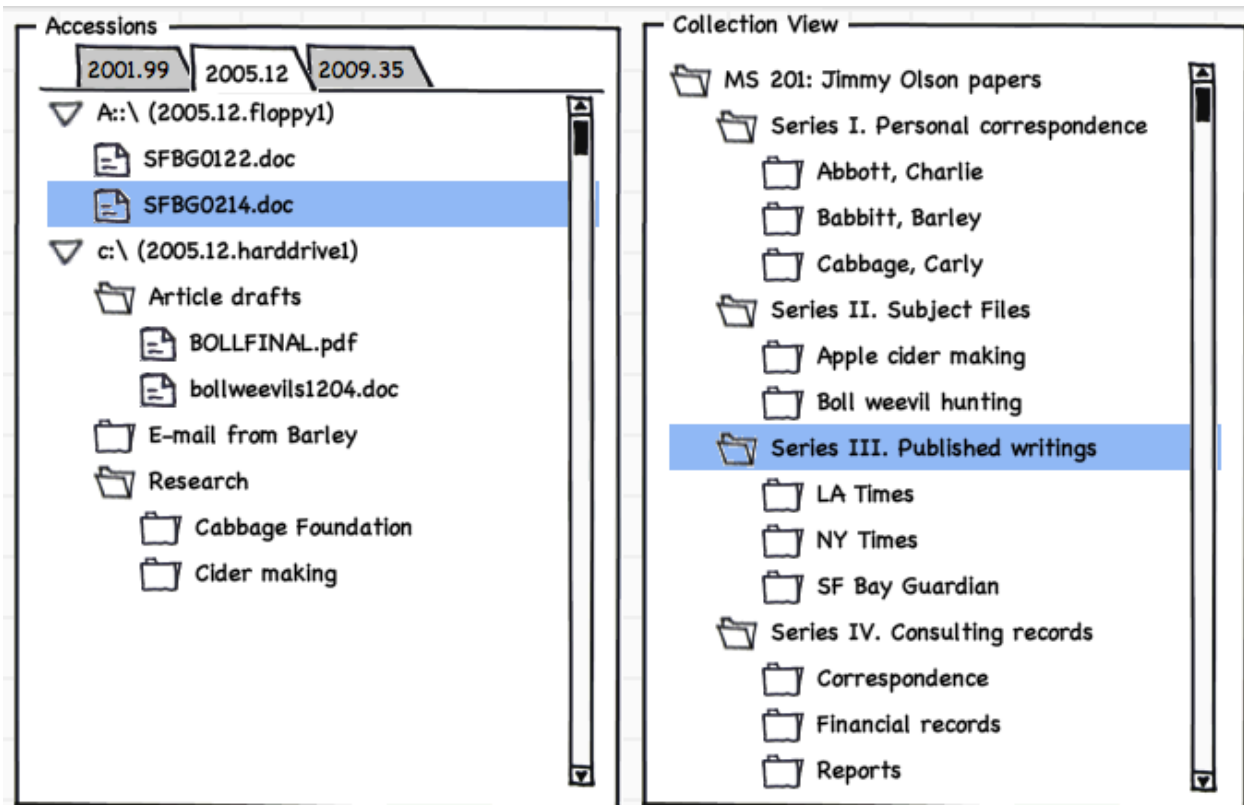
As noted in the Overview above, drag and drop is part of the UI necessary to create an intellectual arrangement for an accession. The original accession is read-only and cannot be modified (with the exception of appraisal actions; see below and **A&D_10**) and represents the original directory structures as they existed within an accession. Dragging a directory, file, or multiple of either to a component in the intellectual arrangement will establish a relationship between those directories and files and that component level.

Component levels also must be draggable to allow for ordering and changing the level of hierarchy. This includes changing the sequence of nodes, promotion and demotion nodes, and auto-renumbering sequences of intellectual units in accord with the modifications.

TomL (Feb 7): Are series ordered by number? Up to this point in the requirements, a programmer would assume “folders” are ordered by the usual rules: date or alphabetic. It is a special requirement that series folders have a numerical sequence.

The UI needs to make it clear which files and folders in the original ingest have or have not been assigned to a component within the intellectual arrangement. Deleting the relationship between a directory or file and the component to which it is assigned should update the status (to “unassigned”) as appropriate.

Following the user interface conventions of desktop file managers, the original ingested accessions could be represented in a pane on the left side of the window, and the intellectual arrangement could be represented in a pane on the right side of the window. Files and folder can be dragged from left to right. The left side should be impossible to modify, with the exception of the ability to remove files during appraisal (see **A&D_10.01** below).



A&D_01.03: Sort records

Archivists will need the ability to sort and filter items within the list of ingested files. Both would apply to these fields: full path, base folder, file, time stamp, size, file type (PRONOM PUID). Ideally, we could apply more than one filter and allow filters to at least have and/or logic against other filters. We will probably need to group PUIDs by larger types: text files, word processing document, HTML, XML, various types of data, etc.

A&D_01.04: Copy and paste of hierarchical structure

*NOTE: See also **A&D_12.05** and **A&D_12.06***

Archivists should be able to copy and paste intellectual arrangement from a number of sources. First, they should be able to copy directory structures from accessions to replicate them in the intellectual arrangement when the directories represent a clearly defined original order. They should also be able to copy existing intellectual arrangements that have been either imported into or created within the tool and paste them into the arrangement pane to duplicate structure as needed.

A&D_02: Technical Metadata (PC)

Overview

The decisions archivists make in terms of appraisal is partly reliant on technical metadata. Technical metadata should

be only viewable and not editable. Technical metadata may also be used to sort records (see A&D_01.03) or in the generation of reports (see A&D_08).

A&D_02.01: File-level technical metadata

The A&D tool should be able to import and provide access (and batch applicable) to the following technical metadata for a given file.

Filename.

Original full file path.

MD5 Hash. The MD5 (16 bytes) hash of the file

SHA-1 Hash. The SHA-1 (20 bytes) hash of the file

File Dates. Lists the Dates and Times of the following activities for that file on the imaged source:

- Created
- Last accessed
- Last modified

File Size.

File Format, as represented by PUID or MIME type. In addition, file format information ideally should have user recognizable names such as WordPerfect 4.2, Lotus 1-2-3 1.2, Word 6.0, etc. and be grouped into the following file categories:

- Archives. Archive files include Email archive files, Zip, Stuffit, Thumbs.db thumbnail graphics, and other archive formats.
- Databases. Database files such as those from MS Access, Lotus Notes NSF, and other database programs.
- Documents. Includes recognized word processing, HTML, WML, XML, TXT, or other document-type files.
- Email. Includes Email messages from Outlook, Outlook Express, AOL, Endoscope, Yahoo, Rethink, Udder, Hotmail, Lotus Notes, and MSN.
- Executables. Includes Win32 executables and DLLs, OS/2, Windows VxD, Windows NT, Java Script, and other executable formats.
- Graphics. Lists files having the standard recognized graphic formats such as .tif, .gif, .jpeg, and .bmp, etc.
- Internet/Chat Files. Lists Microsoft Internet Explorer cache and history indexes.
- Multimedia. Lists .aif, .wav, .asf, and other audio and video files.
- Presentations. Lists multimedia file types such as MS PowerPoint or Corel Presentation files.
- Spreadsheets. Lists spreadsheets from Lotus, Microsoft Excel, QuattroPro, etc.
- Unknown Types. Lists files whose types the A&D tool cannot recognize.

Further Questions or Comments

Categories might need to be configurable for individual institutions.

A&D_02.02: Directory-level technical metadata

If possible, the tool should also provide the following technical metadata at the directory level:

- File Count. The total number of files within a directory.
- Size. Total size of all files in a directory, as expressed in kilobytes, megabytes, gigabytes, etc.
- Creation dates. A range of all files within the directory.

A&D_02.03: Presentation of technical metadata

Users should be able to view the technical metadata presented in a column format that presents the metadata as key/value pairs.

Sample Screenshots

	Name	Path	Item #	Ext	Category	L-Size	MD5	Created	Accessed	Modified
<input type="checkbox"/>	!SAMQ...	/CM05...	8056	ksc	Unknown	38 B	6692E4...	12/31/...	12/31/...	7/6/19...
<input type="checkbox"/>	!SAMQ...	/CM06...	9011	ksc	Unknown	38 B	7699C...	12/31/...	12/31/...	7/6/19...
<input type="checkbox"/>	295.doc	/CM11...	5002	doc	Microso...	49.00 KB	7EF981...	11/30/...	5/1/20...	5/1/20...
<input type="checkbox"/>	296.doc	/CM11...	5004	doc	Microso...	53.00 KB	C52F4...	2/15/2...	5/1/20...	5/1/20...
<input type="checkbox"/>	296.DOC	/CM11...	5024	doc	Unknown	336 B	EE16F...	5/2/20...	12/31/...	5/2/20...
<input type="checkbox"/>	297.doc	/CM11...	5003	doc	Microso...	44.00 KB	08D38...	1/11/2...	5/1/20...	5/1/20...
<input type="checkbox"/>	297.DOC	/CM11...	5023	doc	Unknown	286 B	3D256...	5/2/20...	12/31/...	5/2/20...
<input type="checkbox"/>	299.doc	/CM11...	5005	doc	Microso...	45.50 KB	511174...	4/4/20...	5/1/20...	5/1/20...
<input type="checkbox"/>	2SIDEEVY	/CM06...	75012	<missin...	WordP...	6355 B	9C2A2...	12/31/...	12/31/...	1/5/19...
<input type="checkbox"/>	2SJGBI...	/CM05...	8008	bk1	WordP...	22.82 KB	58E181...	12/31/...	12/31/...	6/15/1...
<input type="checkbox"/>	2SJGBI...	/CM11...	4004	wpd	WordP...	22.73 KB	5D610...	12/31/...	12/31/...	6/20/1...
<input type="checkbox"/>	2SJGBI...	/CM05...	8005	wpd	WordP...	22.73 KB	136F66...	12/31/...	12/31/...	6/15/1...
<input type="checkbox"/>	300.doc	/CM11...	5006	doc	Microso...	42.50 KB	052C5...	5/2/20...	5/1/20...	5/1/20...
<input type="checkbox"/>	301.doc	/CM11...	5007	doc	Microso...	51.00 KB	34B572...	6/13/2...	5/1/20...	5/1/20...
<input type="checkbox"/>	302.doc	/CM11...	5008	doc	Microso...	47.50 KB	940485...	8/1/20...	5/1/20...	5/1/20...
<input type="checkbox"/>	304.doc	/CM11...	5009	doc	Microso...	39.50 KB	0F0060...	10/3/2...	5/1/20...	5/1/20...
<input type="checkbox"/>	9BASE...	/CM08...	2006	<missin...	WordP...	41.56 KB	D16F6...	12/31/...	12/31/...	9/29/1...
<input type="checkbox"/>	A::A:\	/CM07...	1001		Placeh...	n/a		n/a	n/a	n/a
<input type="checkbox"/>	A::A:\	/CM08...	2001		Placeh...	n/a		n/a	n/a	n/a
<input type="checkbox"/>	A::A:\	/CM14...	3001		Placeh...	n/a		n/a	n/a	n/a
<input type="checkbox"/>	A::A:\	/CM11...	4001		Placeh...	n/a		n/a	n/a	n/a

Loaded: 877 | Filtered: 877 | Total: 877 | Highlighted: 0 | Checked: 0 | Total LSize: 41.49 MB

Forensic Toolkit

A&D_03: Descriptive Metadata

Overview

It is essential that this tool is able to create, edit, import and export descriptive metadata about the collection (which might include paper archives not present or represented in Fedora in any way) for use in a third party collection management software (including, but not limited to Archivists Toolkit and CALM). The elements of the descriptive metadata should map to the descriptive elements of Encoded Archival Description (EAD). The tool does not need to store the metadata natively as EAD (e.g., it could store it as MODS), but the tool will need mappings to EAD for both import and export.

The sheer scale of born-digital files means that work is likely to be done over a prolonged period (i.e., over weeks/months). The solutions/workflow must be able to accommodate the flexibility of being able to save work whilst this sorting and processing is on-going.

Further Questions or Comments

It may also mean that more needs to be automated or will be done in less depth - e.g. automating inclusion of subjects / names via “entity extraction” or something like that **or** not doing detailed hierarchies.

A&D_03.01 Importing existing EAD

For hybrid, and multi-accession born-digital collections, there is a strong likelihood that the archival arrangement of the material will already have been undertaken and that the new material will need to be incorporated into the existing structure so that it can be presented as a single collection/finding aid. If material (especially born-digital) is added to a collection, then the existing intellectual arrangement and descriptive metadata must be imported into Hypatia. After importing the existing structure of this collection into Hypatia, the born digital material can be arranged into existing or new series / sub-series etc and then exported as an updated EAD (see *A&D_03.03*).

User Stories

Digital Archivist Carol has just received a deposit of born-digital material from an individual whose paper archives were deposited at the same institution ten years earlier. This additional material from the same depositor will form part of the same archival collection and so Carol would like to import the existing EAD structure into the tool to use as a guide for the arrangement of the born-digital material.

The intern Asok has conducted some initial processing of a new born-digital collection. After reviewing this with the digital archivist, he created the intellectual arrangement that is to be used for this material using AT/CALM. He then exports the entire EAD record which at this time may contain brief details about the accession (i.e., scope/content) and the proposed structure for the collection only - i.e., no descriptive data of the born-digital material.

This information will be used to create the groupings for the intellectual arrangement of the born-digital assets and the foundation for the EAD record. After further work adding descriptive data etc he then exports the updated EAD record so that he can overwrite the version originally created in AT/CALM.

A&D_03.02 Viewing/editing descriptive metadata

The tool will provide the ability to view and edit metadata using a form-based interface. The structure of the collection's intellectual arrangement should be viewable using a tree view (see A&D_01.01). The tool should allow for fields with controlled values (compare with screenshots in A&D_01.01) and allow for both short strings and full text notes for some values.

User Stories

Digital Archivist Tina has imported EAD for a particular collection and uses Hypatia to create an updated intellectual arrangement for the collection. As part of this process it is essential that she is also able to add descriptive data about the series of digital assets that will form the finding aid so needs a data-entry mechanism to add information including title, dates, extent etc. and ideally would like the system to suggest possible content for these fields based upon the items populating that set/folder/series (see **A&D_11**). Tina also needs the ability to assign rights and permissions (see **A&D_04**) at both a folder and individual file level depending on the nature and content of the digital assets.

Further Questions or Comments

Specific clarification is needed for the relationship between the PID for an asset held in Hypatia and its reference in EAD. This could be at least two places in EAD — either the unittid tag or the id attribute on the component levels for the PID associated with the set for a given component (e.g. the series). DAOs should contain references to the PIDs for the files themselves. In addition, the relationship between Hypatia and particular archival data management systems, such as Archivist's Toolkit and CALM, will be needed if users are going to be exporting EAD back and forth between the two systems.

A&D_03.03 Creating new description and intellectual arrangement

For some collections the born-digital material will represent the first accession from that individual/ organization and we must offer the ability to start a completely new intellectual arrangement in the tool rather than force a user to create a skeleton record in AT/CALM and then import it into the tool (as per user story in **A&D_03.01**).

A&D_03.04 Exporting EAD data

Integrating born digital material into an existing arrangement requires that the updated description and arrangement can be successfully re-imported into software such as AT, CALM, and discovery platforms to enable further work or discovery.

Further Questions or Comments

Issues will exist if an institution uses an archival data management system like CALM and Archivist's tool kit. It would be technically very difficult to reconcile an EAD record edited outside of the Hypatia environment with one already ingested, especially if the differences relate to the arrangement of digitized assets. Resolution of these workflow issues are outside of the scope of the tool and will have to be resolved through local practice.

User Story

Digital Archivist Catbert has been working on a hybrid collection for a while and successfully imported the EAD (see *A&D_03.01*) for the paper material and used the Hypatia tool to integrate some born-digital archives. The revised EAD is then exported to CALM and made available to the public as part of the online catalogue.

Two years later a second accession of digital material has been deposited and Catbert then goes through the entire process again by importing the EAD.

Alice is working on a large ingest of born-digital material and having completed the work on one series of born digital assets she now wishes to export the descriptive data into their collection management software so that this material can be made accessible (see *A&D_04*) without having to wait until the entire collection has been processed. She exports the entire EAD back to AT/CALM so that the latest version is held there and can be made discoverable through other procedures. When Alice wants to continue processing the files from this ingest she can re-import the entire EAD from AT/CALM and continue.

A&D_03.05: Controlled vocabularies

Archivists will need to be able to use controlled vocabularies to assign access points at the collection level as well as component levels throughout the intellectual arrangement. The tool should either be able to import existing vocabularies (see *A&D_07.02*) or provide dynamic lookups against existing web services. Additionally, archivists will need to occasionally define new terms (e.g. authorized forms of names that don't currently exist in authority files).

A&D_04: Rights/Restrictions

Restrictions may affect the discovery, retrieval, or delivery of archival material, and will need to exist as controlled values that are machine actionable that have notes for human interpretation.

From the *SAA Glossary of Archival Terminology*: **Access restrictions** may be defined by a period of time or by a class of individual allowed or denied access. . . . **Use restrictions** may limit what can be done with materials, or they may place qualifications on use. For example, an individual may be allowed access to materials but may not have permission or right to copy, quote, or publish those materials, or conditions may be imposed on such use.

In terms of the implementation of this tool, access restrictions are the most critical. Archivists using this tool will need to set both date-based access restrictions and access restrictions based on class of user. They will also need the ability to add notes providing human-readable detail for both access restrictions and use restrictions.

Access restrictions should apply to a given component level and all the related files associated with that component level. Occasionally, related files may have more restrictions than their associated level.

A&D_04.01: Date-based access restrictions with automatic removal

Archivists will need to set date-based access restrictions that will be lifted automatically on a given date.

User Story

Miss Piggy, archivist at the Porcine Institute, is processing the Porky Pig papers. Mr. Pig is a well-known celebrity. The deed of gift for the collection states that two sets of digital records, subject files and correspondence on bacon

addiction, will be restricted only to archivists at the Porcine Institute until 2012. Miss Piggy needs to specify the date-bound access restriction to ensure no one except Porcine Institute archivists will have access to these records. However, Miss Piggy wants researchers to be able to discover these sets of records because they will be open for access soon. Miss Piggy also wants to ensure that the restricted material is available as soon as 2012 begins (i.e., on January 1, 2012).

A&D_04.02: Access restrictions to be removed manually at a later date

Archivists will need to add date-bound access restrictions that cannot be calculated automatically. These will need manual review and presumes that there will be a mechanism to report on restrictions for a given collection (cf. A&D_08).

User Story

Andrew, university archivist at Wilkes-Krier University, is describing the records of the Faculty Committee on Weasel Recovery. This committee discusses student academic issues, and folder titles identify students by name. For FERPA compliance, the records are restricted for the lifetime of the student plus 50 years, or 100 years after the date of creation. Since this restriction cannot be lifted automatically, Andrew wants to add a note describing the restriction as well.

A&D_04.03: Access restrictions for multiple classes of users and individual users

Archivists will need the ability to grant varying levels of access to archivist-defined groups of users and the occasional individual user.

User Story

Andrew (archivist from A&D_04.02 user story) needs to restrict these folder descriptions so only archivists can discover and view them. He needs to ensure that they are not discoverable or viewable by the public, but he may need to grant permission to current committee members or administrative staff at the University on a case by case basis.

A&D_04.04: Variable levels of discovery and access

Archivists will need to have variable levels of gated discovery and access. Levels of access should include “discover” (allowing items to be searched), “view” (allowing metadata to be viewed), “render” (allowing browser-renderable representations of an asset to be displayed), and “download” (allowing associated files to be downloaded).

User stories

Pepe, archivist at Feels Goodman College, needs to set access restrictions on a set of digital records so that they can only be viewed or downloaded from within the FGC Special Collections Reading Room. He wants them to be discoverable, however, and he wants to be able to give individual researchers permission to view them offsite from within their browser. He also needs to add a note describing the on-site use restriction since the records include proprietary software for which FGC has received a special license.

Frank N. Furter is an archivist at the National Organization of Hot-dog And Nitrite-laden Delicious Sausages (NOHANDS). To protect the intellectual property of NOHANDS, he wants to ensure that digital records made available through their discovery and access system are not downloadable. However, he needs researchers to be able to view the browser-renderable versions of the records when they use the system. He wants to set these permissions as he arranges and describes records.

Further Questions or Comments

More nuanced restriction setting may be needed in different situation in the future.

A&D_05: View Files / Representations

Archivists will need to view files or representations of those files to assist in the processes of arrangement and description. The file viewer should have a zooming function. There will obviously be some limitations in providing a viewer for some file types, so alternatives need to be available in some cases. Viewing files should not alter the technical metadata associated with the files, such as access and modification timestamps.

A&D_05.01: View original files

Whenever possible, users should be able to view the original files as rendered in the browser. At a minimum, this should include files that are easily rendered within web browsers (e.g., JPEG, GIF, PNG, text, HTML, PDF, XML, etc.). Ideally, the tool should provide a mechanism render common formats such as Microsoft Word and WordPerfect as well. The viewer should present original formatting whenever possible.

A&D_05.02: Extract and view text strings

For all files (particularly for file formats that are not easily renderable within a web browser) the tool should present a plain text representation of the data within a file by extracting strings.

A&D_05.03: Download files

For all files, archivists should be able to download the files to their local machine to allow them to view them with supplemental software. This can include native software (e.g. versions of WordPerfect) or software that can parse a number of file formats (e.g. QuickView Plus).

A&D_06: Export Metadata (excluding EAD)

Exported metadata formats required:

- METS for an entire collection
- MODS for a single object
- CSV export of all file objects, with associated PIDs/URLs, to be imported into an archival data management system like Archivist's Toolkit.

Questions

Technical metadata could also be exported. The ability to import technical metadata into CALM is a requested feature. However, this type of export is not seen as a priority by the AIMS partners at the moment.

A&D_07: Import Metadata (excluding EAD)

Archivists may want to import descriptive and arrangement metadata from another tool into the arrangement and description tool.

A&D_07.01: Import metadata from Forensic Toolkit

The A&D tool should be able to import the bookmarks, labels and flag “privilege” in collections. Bookmarks will be mapped to series, subseries, etc. Flagged “privilege” items will be mapped to “restricted” materials. Mapping of the “labels” will be decided later.

User Story

Peter, digital archivist at FRED Institute, used the bookmark, label, and flag “privilege” functions in AccessData FTK to assign intellectual arrangement to several collections. There are new accessions to the collections and the A&D tools is available, he wants to import the bookmarks, labels and flag “privilege” he assigned to those collections and process the new accessions using the A&D tools.

A&D_07.02: Import controlled vocabularies

The tool should be able to import controlled vocabulary terms for use within the tool. Sources of data could include Archivists’ Toolkit, CALM, and web services such as id.loc.gov.

User Story

Peter, digital archivist at Present Institute, would like to use subject headings from Archivists’ Toolkit to describe (series / subseries) of the collection he is working on.

A&D_07.03: Import descriptive metadata using entity extraction software/service

The A&D tools should be able to produce entities (name, subject, place) using its an entity extraction engine, third party entity extraction web service, or third party entity extraction program and store the entities extracted in RDF format in Fedora. The entities will become the facets of the collection.

User Story

Peter, digital archivist, at Future Institute, is asked to process a collection with 5 million files. The files are not very organized. He was given 2 weeks to assign descriptive metadata to the files. He expects people using the collection would rely more on full text search and entities (people, places, etc.) browsing but not so much on EAD. He decided to prepare a EAD with very very high level arrangement of the collection and to publish entities extracted using OpenCalais, a very popular entity extraction service.

Screenshot

The screenshot displays the OpenCalais interface for a document titled "Impeaching a Self-Appointed Judge". The interface includes a header with the CALAIS logo and "Powered by Thomson Reuters". Below the header are buttons for "Show RDF" and "Entry Page".

Social Tags:

- Royal Society (☆☆☆)
- Science (☆☆☆)
- Intelligent design (☆☆☆)
- Charles Darwin (☆☆☆)
- On the Origin of Species (☆☆☆)
- Phillip E. Johnson (☆☆☆)
- Darwin on Trial (☆☆☆)
- Discovery Institute campaigns (☆☆☆)
- Religion and science (☆☆☆)
- Pseudoscience (☆☆☆)
- Evolution (☆☆☆)
- Evolutionary biologists (☆☆☆)

Entities:

- City
- Facility
- Industry Term
- Movie
- Organization
- Person
- Position
- Technology

Events & Facts:

- Generic Relations
- Person Career
- Quotation
- Trial

Document Content:

Review of **Darwin on Trial** by **Phillip E. Johnson**, **Washington, D.C.**; Regnery Gateway, 195 pp. 1991.

Stephen Jay Gould

Museum of Comparative Zoology

Harvard University

Cambridge, MA 0238

I teach a course at **Harvard** with philosopher **Robert Nozick** and **lawyer Alan Dershowitz**. We take major issues engaged by each of our professions -- from abortion, to racism to right-to-die -- and we try to explore and integrate our various approaches. We raise many questions and reach no solutions.

Clearly, I believe in this interdisciplinary exercise, and I accept the enlightenment that intelligent outsiders can bring to the puzzles of a discipline. The differences in approach are so fascinating -- and each valid in its own realm. Philosophers will dissect the logic of an argument, an exercise devoid of empirical content, well past the point of glaze over scientific eyes (and here I blame scientists for their parochiality, for all the world's empirics cannot save an argument falsely formulated.) Lawyers face a still different problematic that makes their enterprise even more divergent from science -- and if or two major reasons:

1. The law must reach a decision even when insufficient evidence exists for confident judgment. (Scientists often err in the opposite direction of overcaution, even when the evidence is compelling, if not watertight.) Thus, in capital cases, the law must free a probably guilty man whose malfeasance cannot be proven beyond a doubt (a moral principle that seems admirable to me, but would not work well in science). We operate with probabilities; the law must often traffic in absolutes.
2. There is no "natural law" waiting to be discovered "out there" (pace **Clarence Thomas** in his recent testimony). **Legal systems** are human inventions, based on a history of human thought and practice. Consequently, the law gives decisive weight to the history of its own development -- hence the rule of precedent in deciding cases. Scientists work in an opposite way: we search continually for new signals from nature to invalidate a history of past argument. (As a **sometime historian** of science, I wish that scientists, like lawyers, would pay more attention to, and have more reverence for, their pasts -- but I understand why this is not likely to happen.)

Phillip E. Johnson is a **law professor** at Berkeley, and "a philosophical theist and a Christian" (p. 14) who strongly believes in "a Creator who plays an active role in worldly affairs" (p. 153). Now, I most emphatically do not argue that a **lawyer** shouldn't poke his nose into our domain; nor do I claim that an **attorney** couldn't write a good book about evolution. A **law professor** might well compose a classic about the rhetoric and style of evolutionary discourse; subtlety of argument, after all, is a **lawyer's** business. But, to be useful in this way, a **lawyer** would have to understand and use our norms and rules, or at least tell us where we err in our procedures; **he** cannot simply trot out his criteria and falsely condemn us from a mixture of ignorance and inappropriateness. **Johnson**, unfortunately, has taken the low road in writing a very bad book entitled **Darwin on Trial**.

In a "classic" of anti-evolutionary literature in the last generation, **lawyer Norman Macbeth** (1971) wrote a much better book from the same standpoint, entitled **Darwin Retried** (titles are not subject to copyright). **Macbeth** ultimately failed (though **he** raised some disturbing points along the way) because **he** used an inappropriate legal criterion: the defendant is accused by the scientific establishment and must be acquitted if the faintest shadow of doubt can be raised against Darwinism. (As science is not a discipline that claims to establish certainty, all its conclusions would fall by this inappropriate standard.)

OpenCalais

Social Tags:

Royal Society	☆☆☆
Science	☆☆☆
Intelligent design	☆☆☆
Charles Darwin	☆☆☆
On the Origin of Species	☆☆☆
Phillip E. Johnson	☆☆☆
Darwin on Trial	☆☆☆
Discovery Institute campaigns	☆☆☆
Religion and science	☆☆☆
Pseudoscience	☆☆☆
Evolution	☆☆☆
Evolutionary biologists	☆☆☆

Entities:

City

- Cambridge
- Washington, United States

Facility

- Harvard University
- Museum of Comparative Zoology
- Supreme Court

Industry Term

- all-important reproductive systems
- lateral line systems
- Legal systems
- natural law
- physical systems

Movie

Organization

Person

- Aa Gray
- Abraham Lincoln
- Alan Dershowitz
- Clarence Thomas
- Darwin Retried
- Duane Gish
- Ernst Mayr
- H.F. Osborn
- Julian Huxley
- McInerney
- Norman Macbeth
- Otto Schindewolf
- Phillip E. Johnson
- Robert Nozick
- Stephen Jay Gould

OpenCalais

A&D_08: Reporting

Reporting in an arrangement and description toolset should allow for arbitrary queries. Reports generated from metadata about the records may inform external decision making processes or be used for the calculation of statistics. Reports should be produced in an output format such as CSV or XML that will allow simple post-processing.

A&D_08.01: Report on duplicate items

The tool needs to provide a reporting mechanism that will identify files that have an identical MD5 or SHA1 hash. Because the hash is independent of the filename, identical files may actually have different filenames. (See also: *A&D_01.04, A&D_04, A&D_10.02*)

User Story

Marmaduke, digital archivist at the Great Danish State Library, is processing the Scooby Doo papers. This collection was very disorganized when it arrived and processing the paper component led to the discovery of lots of duplicate material that Marmaduke's supervisor wanted him to remove. Marmaduke wants to create a report of multiple files with identical checksums to help him identify records that can be removed.

A&D_08.02: Report on restricted components and collections

Archivists will need to generate reports listing all collections containing restricted material, as well as all component levels within a specific collection that are restricted.

User Story

Peter Peter, archivist at the Pumpkin Society, wants a report containing all the collections with electronic records that have restricted components. He just needs high-level information. Once he has this information, he discovers that collection MS150, the Gourdie Howe papers, has restrictions. He wants to create another report containing a detailed list of restricted components for MS150 as it is a heavily used collection.

A&D_08.03: Report on file formats

The tool should be able to provide a breakdown of file formats within a collection. This presumes and requires that the technical metadata already is associated with the files. The assumption on our part is that this information is included or generated during ingest.

User Story

Grimace is an archivist for the McDonaldland City Archives. He needs a report containing counts of all the different types of files in RG 12/4/2009, the Mayor McCheese records. He doesn't need to know where each file falls in the collection hierarchy. He also needs an approximate calculation of the total size of the record group. He needs to share this information with H.M. Burglar, the IT director for the City of McDonaldland.

A&D_08.04: Report on appraisal status

Archivists will need to create reports listing the various appraisal statuses as defined in *A&D_10*. There should be both a combined report and separate report for each of the individual statuses.

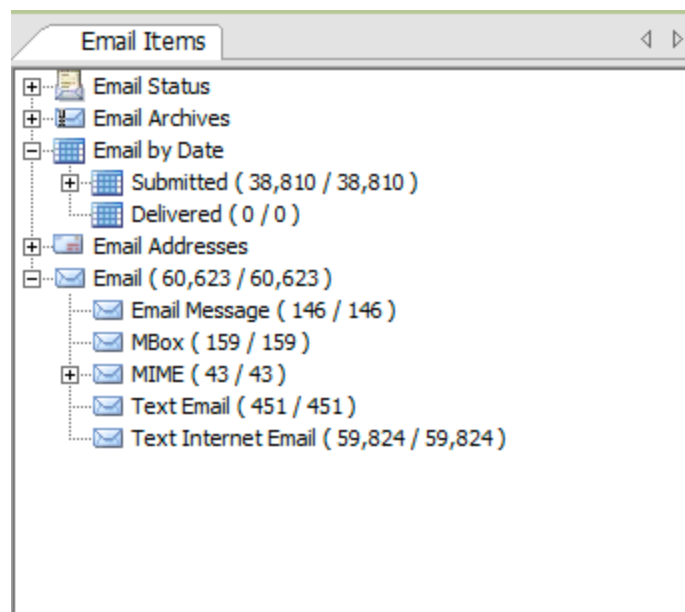
A&D_09: Email

The tool should provide a set of tools to allow work with email messages that may be contained in accessions. The tool should be able to work with email created by different programs (Outlook, Outlook Express, AOL, Yahoo, Hotmail, Lotus Notes, and MSN, Eudora, etc.) and in different formats (mbox, mime, etc.).

A&D_09.01: Display emails by group

The tool should allow archivists to view groupings of emails as follows:

- Email Attachments (Contains only attachments to emails);
- Email Reply (Contains emails with replies);
- Forwarded Email (Contains only emails that have been forwarded);
- From Email (Contains everything derived from an email source, i.e. email related)
- Date (organized by Year, then by Month, then by date, for both Submitted and Delivered);
- Email Addresses (organized by Senders and Recipients, and subcategorized by Email Domain, Display Name, and Email Addresses).

Screenshots

Forensic Toolkit

A&D_09.02: Export/download email

The A&D tools should be able to export emails (cf. *A&D_05.03*) to work with other programs (e.g. network graph, etc.). Ideally, users would be able to select what fields and range of the value of the fields to be exported: to, from, date, cc, bcc, subject, email body.

A&D_10: Appraisal of Material

An archivist will appraise the material to ensure that only items wanted for long term preservation are retained. This is a key professional skill and the approach to this will vary from collection to collection. It may occur either pre or post ingest. Where it occurs after ingest there is a need to record the decision along the same lines as with duplicate files (see *A&D_10.02*). With paper archives we usually ask the depositor whether they want items that we do not wish to retain returned to them, recycled (for non-confidential material) or confidentially destroyed.

A&D_10.01 Marking files for deletion or other appraisal actions

It would be nice for the appraisal process to be able to flag the status of files and folders within the accession ingest as either “keep” “unsure” and “bin”. This should be applicable at any level and inherited downwards but with the ability to change individual file(s) as needed - for example the vast bulk of a series of nested folders should not be kept but there are a few individual files that should be retained (or vice versa)

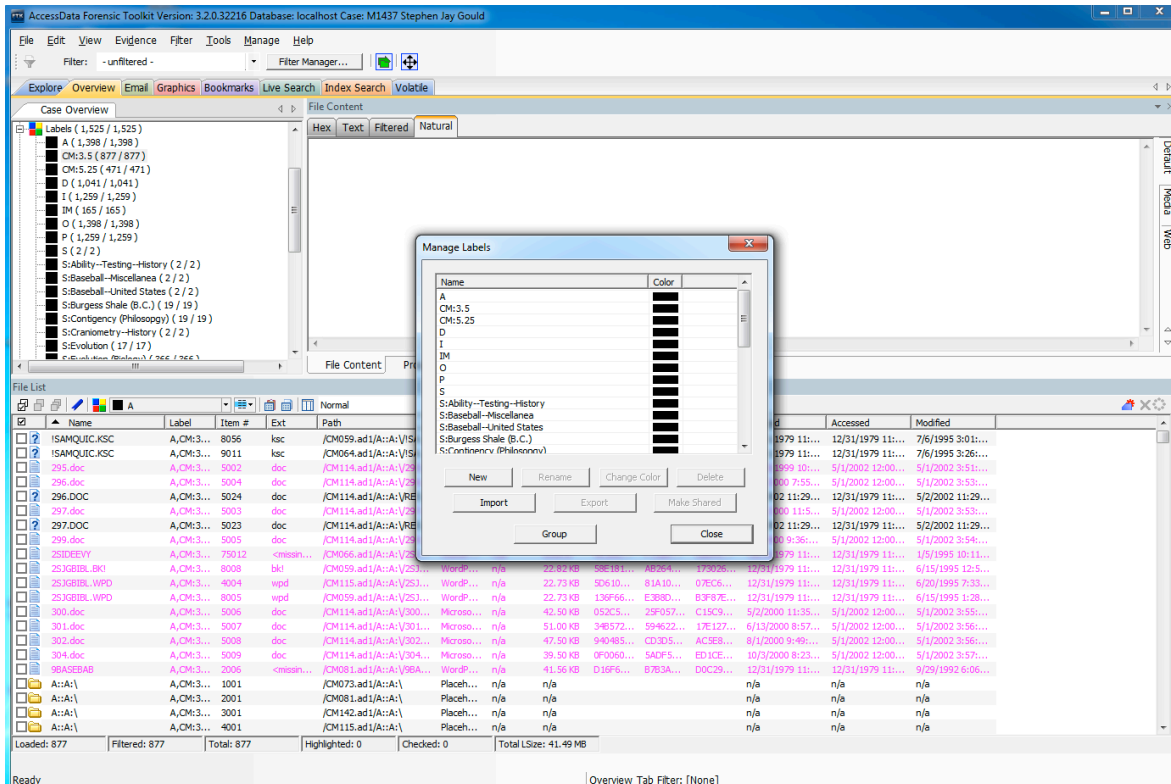
With large accessions it might be that similar material is held in different folders - so the ability to sort or filter the files (see *A&D_01.03*) in the accession ingest could offer the flexibility of looking at the material in alternative ways - but this should be a **temporary situation** and should not change or over-ride the arrangement of the folders/files at the point of ingest.

It is likely that appraisal will be conducted over time so the appraisal status flag would assist with recording progress through the material or allow additional staff to review particular sections of material (i.e., everything marked unsure).

In many situations it will not be possible to make an appraisal decision based purely on the technical metadata that is available so the archivist would need to open a file to make the decision about whether it is kept or not. It is essential that this appraisal process does not impact upon the technical metadata (especially last opened/accessed) date that we subsequently wish to present to researchers. Therefore, two options to achieve this might be to generate an access copy at the point of ingest or to ensure that “last modification time” for a file is the last mod time the file had when it was originally ingested and will not be overwritten if the file is opened.

For material to be deleted there should be a two-step process requiring confirmation etc. — one option could be to get Hypatia to generate a report listing all of the files to be deleted but I am not sure how useful/practical this would be if hundreds (or more) files are being deleted. We should consider making deletion of files from the ingest something that is restricted to particular user roles with appropriate permissions (ref to **A&D_04**). The return or destruction of the born-digital files may have been indicated during the deposit/transfer process. This work will be done outside Hypatia. A note regarding the removal of material as part of the appraisal process, such as a broad note like “third-party publications removed” should be possible any component level and at the collection level. It should correspond with the EAD note element <appraisal/>.

Screenshots



Forensic Tolokit (application of labels)

A&D_10.02: Duplicate Files

Either as part of the appraisal process or otherwise there is a requirement to be able to detect files that are exact duplicates of another file in the repository and to then be able to either to hide or delete the file. It is critical that all actions on the file are automatically record to provide a full audit trail.

User Stories

Digital Archivist Dilbert, based at the Scott Adams University, likes to keep a tidy ship and knows that this includes the digital repository and hates the thought of storing, preserving and providing access to multiple versions of the same digital file - whether this is because of a user accidentally misfiling a file into a specific folder or because the file has been transferred as part of multiple ingests over time. He does know that they are exactly the same because he has asked the processing archivist Wally to run a report (see A&D_08) using their checksum value.

Having run the report to detect duplicate files within a single accession, a single collection (i.e., multiple accessions) or across everything, Wally can look at the report data (this should

include ingest ref?, creation date, last viewed data, filepath and/or location of the matching file(s) and then has three options.

- Hide: This hides a file from view so it is not visible for the archival arrangement, discovery or access elements of the workflow
- Delete: This marks the file(s) as ready for deletion but suggest a further prompt to confirm that you want to delete the file from the system completely. This technique could also be applied to files/folders that are identified for deletion as part of an appraisal process and/or files that subsequently need to be removed (e.g., for copyright purposes) [should Wally be able to delete files?]
- Ignore: This says I know that the files are the same but do not wish to hide or delete it

For the purposes of creating the report and accessing the audit trail etc all hidden and deleted files need to have a datastream updated to reflect the change with possibly a default content - *this file was identified as a duplicate by XXperson on YYdate or deleted by XXperson for ZZ reasons.*

Screenshots

Accessed	Modified	Flagge...	Duplica...
12/31/...	7/6/19...	False	
12/31/...	7/6/19...	False	
12/31/...	4/22/1...	False	
12/31/...	4/22/1...	False	
12/31/...	4/22/1...	False	
12/31/...	4/22/1...	False	
12/31/...	4/22/1...	False	
12/31/...	4/22/1...	False	
12/31/...	4/22/1...	False	
12/31/...	4/22/1...	False	
5/1/20...	5/1/20...	False	
5/1/20...	5/1/20...	False	
12/31/...	5/2/20...	False	
5/1/20...	5/1/20...	False	
12/31/...	5/2/20...	False	
5/1/20...	5/1/20...	False	
12/31/...	1/5/19...	False	Primary
12/31/...	1/5/19...	False	Second...
12/31/...	6/15/1...	False	
12/31/...	6/20/1...	False	
12/31/...	6/15/1...	False	

Forensic Toolkit

Further Questions or Comments

Archivists should be able to decide which duplicate file should be the primary. Technical metadata could be used to determine the oldest duplicate and make that the primary.

When deleting files, it is unclear if Hypatia should delete the files itself or just “identify” them for deletion. If we choose the latter, this would mean that the deletion would happen outside of Hypatia.

When you delete a file (or series of files) should we delete the physical file including any derived versions etc. but leave a “shadow” or “tombstone” record that includes an audit trail and reason for deletion (i.e. duplicate or appraisal). This should be available as distinct report (see *A&D_08*). In addition, the issue of whether or not preservation copies of “hidden” files or files marked as “deleted” but not removed should be created. This may be out of scope for Hypatia, but local implementation should consider the issue.

A&D_10.03: Immediate (unstaged) deletion

The tool should provide an option to delete files immediately if needed. This must present a confirmation screen as files may not be recoverable. This functionality should still retain a “tombstone” record that includes the date of deletion.

A&D_11: Batch Application of metadata from files

The sheer volume of files means that we should try to automatically use the extractable metadata to form the proposed basis of the descriptive metadata. For example:

A&D_11.01: Apply filename to title field

A&D_11.02: Apply creation/modified date to “from” date field

A&D_11.03: Apply access/modified date to “to” date field

A&D_11.04: Apply creator to creator field

A&D_11.05: Apply file format to descriptive/technical note

A&D_11.06: Apply number of files to extent

A&D_11.07: Apply size of file(s) to extent

Further Questions or Comments

When multiple assets are being described by a single descriptive entry, the information could be derived from the first file it encountered (sorted by name or date etc) or from the directory level, see *A&D_02.02*.

The ability to define the date format to be used would be a “nice” feature given US vs UK local practices. However, the date will probably be stored in a machine readable format which will allow us to easily customize how it gets presented to the end user.

A&D_12: Intellectual arrangement

The contents of this overview have been adapted from section D2, “Resources in AT Description,” in the functional requirements for the Archivists’ Toolkit Description Module (<http://archiviststoolkit.org/sites/default/files/description.pdf>, pp. 4-9).

Record Types

A **collection** is 1) an item or aggregate of items generated or collected by an individual, family, or organization in the course of their activities and deemed to be of enduring value, 2) and is in the custody of an archival institution.

Collections may also be linked to **components** to form multi-level descriptions.

Hierarchical Levels

These two record types and their associated interfaces for descriptive metadata (see **A&D_03**) accommodate the twelve levels of description permitted in the Encoded Archival Description standard. In other words, a **collection** in the A&D tool may be represented by up to twelve hierarchical levels of records. A **collection** record may be the parent of a **component** record that is parent to a **component** record that is parent to a **component** record, and so on up to twelve levels deep. There may be an unlimited number of component records at each level, that is, there is no limit on the number of series records or file records. Records at the same level are referred to as **sibling records**.

EAD provides a standardized vocabulary of labels for the permitted hierarchical levels in an archival resource. These labels (**class, collection, file, fonds, item, otherlevel, record group, series, subfonds, subgroup, subseries**) each correspond, or map, to one or more of the collection or component records (See Table 1 in AT Description specification).

When the operator chooses to add a new component record to an existing collection or a component record, she or he **must choose** a level label for the component. The options given the operator are driven by a set of rules for acceptable children for a given level. For example, the parent of a subseries can be a series, but not a collection. (See Table 2 in AT Description specification)

Intellectual and Physical Order of Archival Resources

Intellectual hierarchy will be captured by **tracking the relationship** of the collection records and component records to each other. **Both parent/child record linkages and sibling record sequences must be captured and stored.**

A&D_12.01: Create new collections

Archivists must be able to create new records representing archival collections. Descriptive metadata should follow the collection-level elements available within Encoded Archival Description and **must include** creator, title, date ranges, identifiers, and call numbers.

User Story

Eugene is processing the Absurd Theater Records. The collection does not have an existing EAD finding aid or description in another system. He loads up the tool and logs in to his account. Once logged in, he selects “Create new collection.” He enters the metadata about the collection and clicks save. Once he saves, he is redirected to a page for that collection.

A&D_12.02: Create new component levels

Archivists must be able to create new component levels that are children of collection records or siblings or children of other component levels. See **A&D_03** for description-related requirements.

User Story

Eugene then needs to create a new series of subject folders in the collection. He is logged into the system and is viewing the collection page. He selects “Create new component.” In the “Level” field, he selects “Series.” He fills out the metadata about the series and clicks save. Once he saves, he is redirected to page for that series.

A&D_12.03: Associate files and directories to component levels

Archivists must be able to associate files and directories from accessions with component levels. They may also need to remove or change the associations. Assigning a directory to a component **should not** create a new component within the intellectual arrangement.

The tool should allow for multiple associations during the arrangement process. **However, a file or directory must have relationships with no more than one component within a “finalized” intellectual arrangement.** This reflects constraints on arrangement as defined in archival practice.

A&D_12.04: Associate accession with collection

There will be cases where accessions will not include metadata that relates them to a specific collection, so the tool will need to provide the ability to allow archivists to associate accessions with collections.

User Story

Eugene wants to take files from an accession that have been ingested and assign them to this series. He associates the accession with this collection by selecting the appropriate accessions by number. He also can see a list of all unassociated accessions.

A&D_12.05: Replicate directory structure from accession into intellectual arrangement

Archivists may discover that an accession’s directory structure demonstrates that a creator had a clear existing arrangement that should be maintained. Accordingly, archivists using the tool should be able to replicate some or all of the the directory structure from an accession into corresponding component levels. See also **A&D_01.04**, **A&D_02**, and **A&D_11**.

A&D_12.06: Duplicate components and structure in intellectual arrangement

Archivists are used to being able to copy component structure during the arrangement process to prototype various intellectual arrangements. Contents of the descriptive metadata should be duplicated automatically.

A&D_13: Searching within files

Ability to do pattern or keyword searches in order to discover files that should be restricted - credit card or social security information; passwords; student or medical files etc.

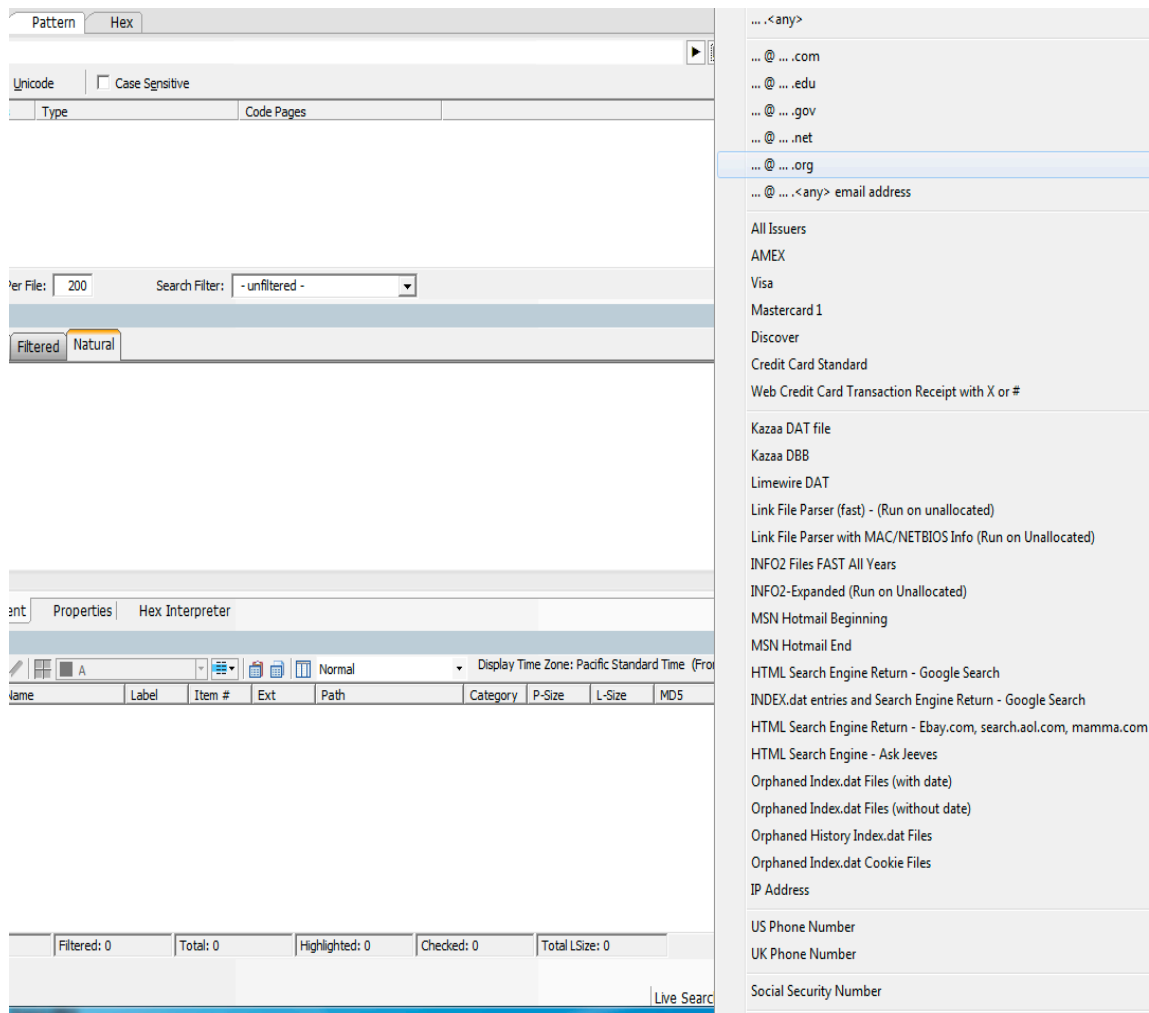
A&D_13.01: Pattern searching

For pattern search, it is desirable to allow users to define their own patterns as well as to include commonly used patterns such as social security number; phone no., credit card nos. etc.

User Story

Peter, digital archivist at FRED Institute, is processing a born digital collection. He is concern on the existence of social security numbers in the files. He would like to perform a search on the whole collection so that files containing texts with XXX-XX-XXXX (X- numeric) pattern will be grouped with the text XXX-XX-XXXX highlighted for him to review for restriction.

Screenshot



Forensic Toolkit (Pattern Search)

A&D_13.02: Full-text searching

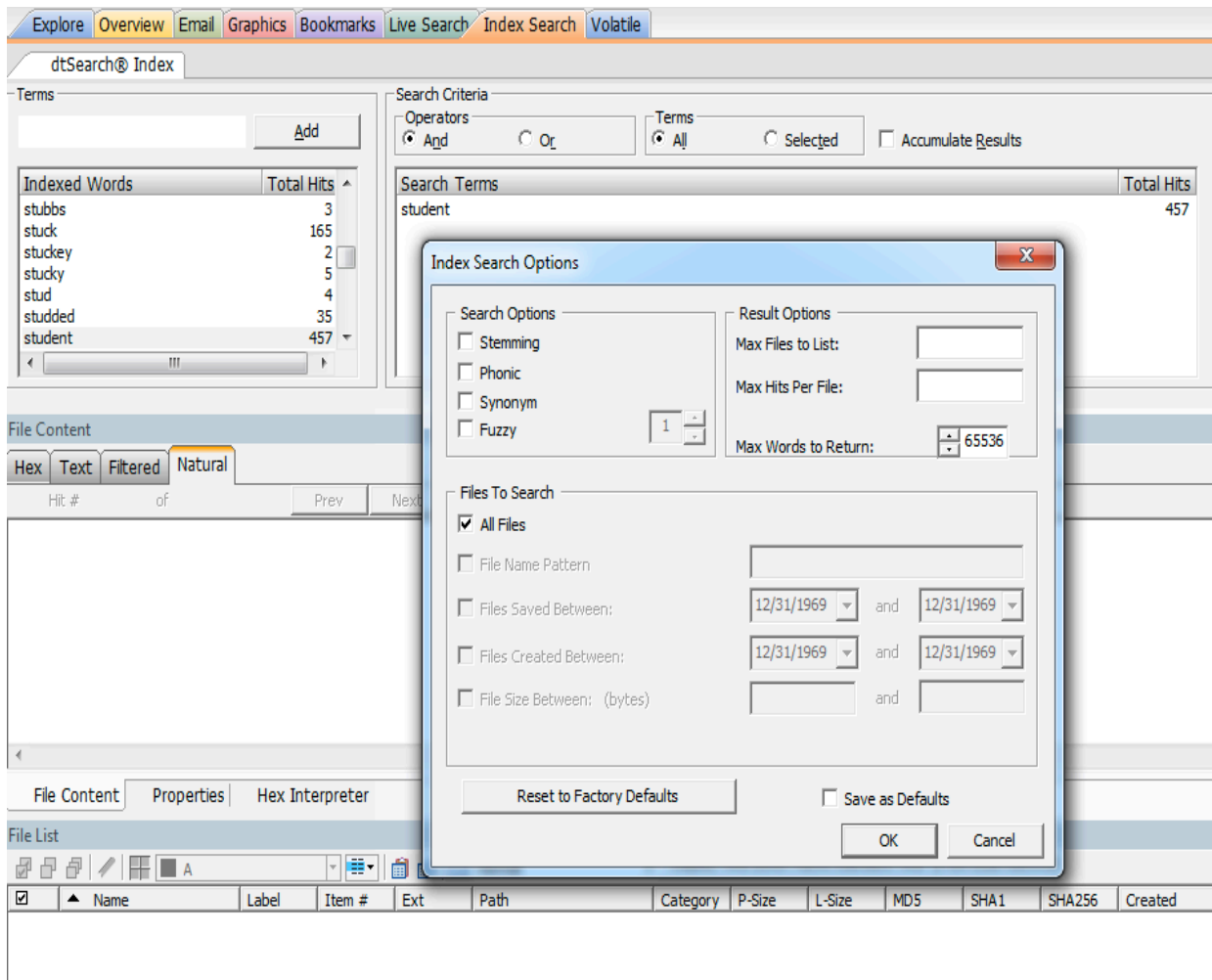
For full text search, it is desirable to have the following options:

- Stemming. Words that contain the same root, such as raise and raising.
- Phonic. Words that sound the same, such as raise and raze.
- Synonym. Words that have similar meanings, such as raise and lift.
- Fuzzy. Words that have similar spellings, such as raise and raize

User Story

Peter, digital archivist at FRED Institute, is processing a born digital collection. He was told by his supervisor that all files containing student grades should be restricted. He would like to perform search on the whole collection so that all files with “student”, “students”, “grade”, “grades” will be grouped and the texts “student”, “students”, “grade”, “grades” highlighted for him to review for restriction.

Screenshot



Forensic Toolkit (Full-text search)

Addendum: Functional Requirements Based on Yale Workflow Diagram (For Discussion)

Prepare for Arrangement

Use accession / acquisition records, existing surveys and descriptions, inventories, biographies, donor documents / correspondence to prepare for arrangement.

Component Tasks

- Select collection for processing
- Gather records and information
- Assign to processing archivist
- Restrict collection during processing

Survey Collection

Use appropriate tools to survey the records.

Component Tasks

- Assess / analyze type / condition of the media
- Assess / analyze file formats, sizes, dates
- Assess / analyze existing arrangement (e.g. folders)
- Assess / analyze content
- Assess / analyze context, functions

Arrange records intellectually

The intellectual arrangement of archives is a critical process in making the material, regardless of format, accessible to users. Wherever possible/practical the skills and terminology applied to paper materials should be applicable to born-digital materials. Archival collections are usually catalogued according to ISAD(G) or DACS cataloguing standards which are non-format specific.

Archival arrangement is a key professional skill, with a user likely to make a number of assumptions about the material depending upon its intellectual arrangement. Working with born-digital material is likely to be harder than working with paper records due to the increased practice of mixing both original and material from third-parties into a single filing system and the much greater volume of material concerned. For organizational records there is the increased complexity as a result of individual, team and institutional file stores.

Digital archives offers the potential to place a single digital asset in multiple locations with-in an archival arrangement and we should resist this temptation and retain the current practice of a single unique place with-in the intellectual arrangement and to include appropriate cross-references to aid users.

Component Tasks

- Review the material including its context and content and see if it is possible to identify or determine the “original order” for the material
- Create a logical order if the original order cannot be identified
- Identify the processing and cataloguing requirements for this particular collection [Note this could be part of the survey stage?]
- Place material of similar nature (e.g., all files relating to Book X) or function (e.g., all minutes of a specific committee) into series
- Create a hierarchy of the material into cascading series' from Collection at the top to Item (a single digital asset) at the bottom [the sheer volume of digital material means that the predominant practice is likely to be cataloguing at series level]
- If material is already held, the deposit of additional materials (whether paper or born-digital) will need to be integrated into the existing intellectual arrangement

Note: Heavily linked with issues of the GUI (**A&D_01**), import and export of EAD (**A&D_03**) and feeds into Descriptive Metadata

Further Questions or Comments

Consideration of access and permission issues (see **A&D_04**) should be done at the highest level first - e.g., apply the conditions to the collection and then modify specific series/items that vary from this position (e.g., a collection may be generally open but a specific series of records closed for xx years under Data Protection legislation)

Arrange records physically

With paper records an important element in their management is that relating to its location to enable easy retrieval from the store by the archives staff. For born-digital materials the original file will be ingested into the repository and preserved and an “access copy” version derived from the original created for individuals to access and use.

Whilst the “location” of the original file will remain in the Repository there is a need to create a link so that individuals with appropriate access and permissions can retrieve the access copy digital asset without further involvement by the archives staff. It is important that this link remains persistent for authenticity and citation purposes. It must also avoid revealing the true path of the Repository and so risk unauthorised access to other digital assets.

Component Tasks

- For retrospective cataloguing there will need to be a systematic process of identifying born-digital material within existing material
- Removing these file(s) for processing and subsequent ingest into the Repository
- associating ingested files/folders with Fedora sets to place it into its place in the hierarchy / intellectual arrangement

Create descriptive tools

Use information about content, context, physical characteristics, and archival management processes to create standardized or customized information products for various purposes. A description tool should be able to import (for existing or hybrid collections) and export EAD files. At a minimum, the tool should be able to export the structure of the container list. [Note: see A&D_03 for similar functions]

Component Tasks

- Describe biography/history
- Create scope notes and component description
- Create intellectual structure/box list
- Summarize preservation and appraisal actions
- Create subject and name access

Further Questions or Comments

Defining cataloguing standards for digital materials out is out of scope for the requirements, but the AIMS project at large may want to comment on this.

What level of descriptive activity is necessary for an AIMS-specific description tool?

Is assigning subject terms for different descriptive levels necessary?

What about biographical/historical notes?

In other words, is the AIMS-specific tool narrowly scoped, with the description then exported to something like the AT/CALM for further work?

Perform physical control

Assign archival records to containers and storage locations appropriate to their physical composition, technical characteristics, extent, and condition. Pack, label, and store materials so they can be retrieved and moved as needed. Assign identifiers to groups or containers of archival records. Storage assignments follow plans that reflect the archival institution's policies governing the placement of various materials within facilities.

Component Tasks

- Assign call numbers, locators, barcodes, and other identifiers
- Label boxes, folders, etc. [OUT OF SCOPE]
- Create and updates holdings and location records.
- Send materials to storage location [OUT OF SCOPE]

Further Questions or Comments

We won't really be "assigning" storage locations per se. True storage management is going to be out of scope. What will be needed is 1) basic workflow management that can represent a "commitment" action that work is

completed for a given collection or set of records, 2) a tool that may allow us to add mnemonic identifiers (e.g. based on call numbers), and 3) exposure of PID/URLs in the interface to allow linking back from descriptive tools. There is a larger integration with AT (and possibly CALM) bulk importing digital object locations that probably needs to be hashed out at some point.

Disseminate descriptive tools and records

Prepare records and descriptive tools for dissemination in access systems. Publication and indexing of descriptive tools. Digitization/transcription as appropriate. Creation of dissemination packages of records as appropriate.

Component tasks

- Publish descriptive tools
- Publish records
- Index descriptive tools
- Index records [LOW PRIORITY?]

Further Questions or Comments

This again is largely about workflow and “promoting” our descriptions etc. to a discovery and access tool. We will need a way to signal that these are ready to be discoverable. A way to build formally defined dissemination packages is not needed within the A&D tool, but discovery and access requirements suggest the need for ways of traversing the object relationships that would allow, for example, someone to retrieve all records from a given series, or perhaps all records in a collection. Indexing descriptive tools really falls under the domain of an access and discovery tool.

Complete processing

Signal that records are ready for access. Receive final approval from supervisors and document completion.

Component Tasks

- Remove processing restrictions [OUT OF SCOPE?]

Out of Scope

Requirements that were discussed but deemed out of scope are included here with reference to the section from where it was originally proposed in the document.

From A&D_03.03 Creating new description and intellectual arrangement

Can the information about the donor, deposit etc from the Donor Survey as the basis of an accession-type entry be accessed? Clearly there is a high chance that the potential material identified in the survey will not reflect the actual material subsequently transferred.

How to do this is an issue. Copy and paste would be easy to implement, but painfully slow. Drag and drop would be visually nice, but also painfully slow for more than a few fields. Various aspects of A&D would benefit from a

scripted approach instead of a visual UI. The “scripts” would be akin to macros. Historically, this type of functionality is reasonably easy to implement, and add a huge amount of power and flexibility to a product.

We would also already have access to this information already in multiple ways. The EAD finding aid is the public view, but in AT/CALM which are collection management systems there is also a host of other data you need to record/manage but not divulge to the public within the accession/depositor tables of AT/CALM. e.g., depositor contact details, terms of deposit.

This information is out of scope for Hypatia unless it is relevant to how we arrange material or it's important enough to include in description. However, it should still be considered as it relates to information flow between Hypatia (for discovery, access, and management of digital objects) and AT/CALM for larger archival management. We need to illustrate it has not been forgotten/ignored especially as some of this information may be captured via the web survey

From A&D_03.02 Viewing/editing descriptive metadata

The “master” version of the EAD file should be dictated by local practice when using files created in AT/CALM and subsequently edited in Hypatia.

From A&D_04.04: Variable levels of discovery and access

Tools like FTK allow for restriction at the individual file level. Questions still remain as to whether it is good practice to allow mixing unrestricted files and restricted files into a specific level of arrangement, but I added this as an option above.

Functional requirements from Yale Workflow Diagram - Complete Processing

- Receive final approval [WORKFLOW]
- Document completion
- Announce availability of records and descriptive tools [OUT OF SCOPE]

If we have a project archivist working, we may want to have a senior archivist review work before we mark it as “done” but this is really a question of workflow. “Document completion” should be about generating documentation about what was done in processing, and clearly relates to A&D_08 above. It is unclear what form a processing report would take in this case, or whether it would be important enough to create.

2. Rubymatica

Rubymatica is an open source software project written in Ruby and adapts some of the convenient workflow provided by Archivemata. It is primarily an application programming interface (API) with the purpose of creating an arrangement of files for ingest. The project contains a simple demonstration web site which is open to the public on a by-request basis. Rubymatica adapts some aspects of the SIP to AIP transformation phase of Archivemata as a means to build SIPs ready for ingest.

There were several reasons to create a Ruby version of Archivemata. Rubymatica is written in Ruby so that it can easily be integrated into Hypatia. Ruby has become prevalent for developing web applications, and the University of Virginia (UVA) has standardized on Ruby and Java. At UVA, legacy Python, Perl, PHP web applications are being superseded or converted to Ruby. Writing the tool in Ruby also offered the opportunity to create some additional functionality than what is present in Archivemata.

Rubymatica, being a program to prepare files for Hypatia ingest, has somewhat different goals than the SIP ingest phase of Archivemata. Because of this, there is some different logging, and the creation of metadata databases that aren't necessary in Archivemata. The workflow and general architecture are similar to Archivemata. In both Rubymatica and Archivemata, many of the same external applications handle tasks such as unpacking archives, generating checksums, and checking files for malware. Ruby scripts do the bookkeeping, workflow, and data management. Each external application processes files without any knowledge of the overall workflow. Rubymatica has both command line and web interfaces.

This work only took a few days to complete and as part of the development process, members of the UVA Library software team did a code review of Rubymatica for both legibility and security issues. Rubymatica is on Github along with extensive, technical documentation.

<https://github.com/twl8n/Rubymatica>

Rubymatica processes each ingest as a single process, copying and transforming the ingest into a new directory tree containing a subdirectory with same structure as the original, plus metadata subdirectories. The ingest may be in the form of an archive file (ZIP, tar, or rar files) or a directory. Rubymatica has additional functions to create a BagIt bag, to integrate a Tufts TAPER submission agreement, and to integrate a donor survey. The current version also has a feature to create categories for PRONOM file identifications. PRONOM's DROID application is run via the FITS file identification suite.

Rubymatica runs several applications on every file in a logical copy of an ingest. The processing steps are:

1. Copy original files into a working directory tree
2. Recursively unpack any archive files
3. Cleanse (detox) file names of characters not supported by MS Windows, MacOS, and Linux
4. Check for malware, create checksums, identify file types via FITS and DROID, and write a METS file

Log files are maintained, and several very small databases are created to track metadata and status of the ingest. After this processing, the collection is in a form suitable for assessment and eventual ingest into a repository for further processing.

Archive files are unpacked into new, uniquely named directories in order to avoid directory name and file name conflicts. File name conflicts are also avoided during file name cleansing. Processing happens as a background process in order to prevent a timeout. The background process can (in theory) run as long as necessary to process an ingest.

3. Hypatia

Hypatia is an initiative to create a Hydra application (Fedora, Hydra, Solr, Blacklight) that supports the accessioning, arrangement / description, delivery and long-term preservation of born digital archival collections. Hypatia is being developed as part of the AIMS Project ("Born-Digital Collections: An Inter-Institutional Model for Stewardship"), funded by the Andrew W. Mellon Foundation.

Hypatia is a cross-institutional effort that includes University of Virginia (grant lead), University of Hull, Stanford (Hypatia development lead), Yale, and a third party software development company called MediaShelf.

Functional Requirements for Application Development

At the beginning of 2011 the AIMS digital archivists' created functional requirements for the application. These functional requirements are primarily focused on how an archivist would arrange and describe born digital collection materials in a browser based software application. The functional requirements were used to develop technical development tasks that have been translated into tickets in an Hypatia JIRA project (<https://jira.duraspace.org/browse/HYPAT>). At the end of the current development cycle only a partial set of these requirements will be supported by the Hypatia application. Complete implementation of these requirements will not be complete by the end of the current development effort.

Current Status of Hypatia development

The current phase of Hypatia development will be completed at the end of the October 31, 2011. Hypatia is being developed using an Agile methodology with weeklong iterations and weekly code submissions. The Hypatia application will not be completely functional at the end of this grant cycle and it is anticipated that the institutions supported by the current grant will seek additional funding to continue developing the application. By October 31st Hypatia will have the following functionality:

- A demonstration application hosted by Stanford that contains records for all of the AIMS collections
- A polished interface that allows for the discovery and display of AIMS born digital collections
- A small subset of the AIMS collections will also contain descriptive metadata and digital objects from these collections. All of the content loaded into the demonstration application will be viewable by the public.
- The ability to download disk images and file level assets.
- The ability to create groupings of digital objects (sets).
- The ability to edit descriptive and technical metadata for collections, sets and digital objects.
- Drag and drop functionality to assist archivists in the arranging and describing of born digital collection materials.



Additional information on the Hypatia application can be found at:

Hypatia Project Wiki:

<https://wiki.duraspace.org/display/HYPAT/Home>

JIRA project

<https://jira.duraspace.org/browse/HYPAT>

The Hypatia demonstration application is hosted at Stanford and publically available at:

<http://hypatia-demo.stanford.edu/>



Appendix I: Digital Archivist Community

I. Born Digital Archives Blog

Background

The blog <http://born-digital-archives.blogspot.com/> was created in late May 2009 as part of the digital archivist community building work. The Digital Archivists felt that a blog would offer an easier and quicker mechanism for the digital archivists and software developer to provide updates on their work than placing all of this on the project website hosted by UVA.

It was also a reflection of the usefulness of some digital preservation blogs that we were reading on a regular basis including Chris Prom's Practical E-Records (<http://e-records.chrisprom.com/>) and the FutureArch blog (<http://futurearchives.blogspot.com/>).

Content

A wide range of topics have been featured on the blog including digital forensics, reports of events attended including the DLF Forum, the AIMS un-conference and the 2011 Personal Digital Archiving Conference. It has featured the use of FTK at Stanford, the development of Rubymatica by the project's software developer, the creation of a web survey to collate information from donors and arrangement and description of born-digital archives.

Statistics

As of 24th October 2011:

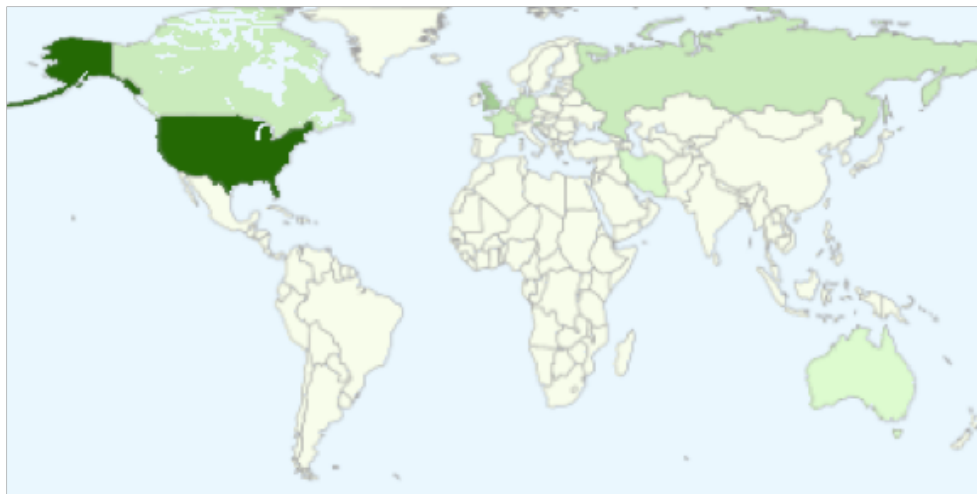
Total Number of posts:	39
Total Number of page views:	19,418
About Us page - number of page views:	1006

Three most popular posts are:

Other Highlights from the DLF Fall Forum (uploaded 16 Nov 2010)	1971 page views
Digital Library Federation (DLF), Fall Forum, 2010 (uploaded 29 Oct 2010)	1236 page views
AIMS; the Unconference (uploaded 18 May 2011)	392 page views

Audience:

Country	Page Views	Percent of Total Views
United States	8466	44%
United Kingdom	2577	13%
France	1199	6%
Russia	945	5%
Canada	920	5%



Impact

It was not practical to post as frequently as we had initially hoped and believed that the postings needed to be relevant and interesting rather than regular. In some cases the nature of the work meant it was not always



appropriate to write an entry – for example the AIMS unconference and UK symposium were largely by personal invitation which removed the necessity to use the blog to generate interest prior to the event.

Future

The four institutions have agreed to continue providing updates of activities and reports of events attended beyond the life of the grant.



2. Digital Archivist Community Events

Part of the AIMS approach has been to situate the framework within the standards and best practices set by the archival community. However, as the project began the partners realized that the community surrounding the specific issues of born-digital materials in collecting repositories was emerging somewhat differently between the US and the UK.

The US had a well-established electronic records and digital preservation community at the outset, but its connection to collecting repositories was not very strong. That said, early efforts must be acknowledged here, such as Susan E. Davis's 2008 article, "Electronic Records Planning in 'Collecting' Repositories" (*American Archivist* 71, no. 1), Michael Forstrom's 2009 article, "Managing Electronic Records in Manuscript Collections: A Case Study from the Beinecke Rare Book and Manuscript Library" (*American Archivist* 72), and the March 2009 Stewardship of E-Manuscripts symposium held at the University of North Carolina. There were relatively few posts with the explicit job title of digital archivist, and the precise requirements and responsibilities of these posts varied quite dramatically. Mark A. Matienzo, Digital Archivist at Yale University, expressed a significant interest in bringing these communities together more frequently. In the UK there was already quite an established digital preservation community with much of the momentum being created by the Digital Curation Centre and the Digital Preservation Coalition. There are however, only a few examples of posts with the explicit job title of digital archivist.

A key element of the AIMS Project was active engagement with these emerging communities, both in order to gather insight and information for the development of the whitepaper, but also to ensure that the framework would have a community in which it could be adopted. To accomplish this, the Digital Archivists participated in different archival and born-digital community events and the AIMS team coordinated three additional outreach events. A complete list of events attended or held by the AIMS team is included at the end of this appendix. A summary of the AIMS-sponsored events follows.

AIMS Unconference, Charlottesville, May 2011

Feedback from AIMS Unconference Attendees:

I most enjoyed the chance to talk to known colleagues and meet new ones. It's also useful to hear what other folks consider a "solved" problem in their environments (and therefore a potentially replicable solution), and what is still completely challenging. For example everyone's recognition that managing access to restricted materials is not supported by current tools was fortifying. I believe that consensus like that is very important for funding agencies to hear, so they can focus on funding projects that aim to chip away at this problem.

- Aprille McKay, University of Michigan

I think this was the perfect professional development activity for me right now. This group was neither too large nor too small and yet specialized enough that we are able to immediately get to the specific issues facing our community. I sometimes feel depleted as the "born-digital" person in my institution - this invigorated my drive and inspired me with new approaches and fresh ideas to get to work on some daunting tasks back home.

- Erin O'Meara, UNC Chapel Hill

The lightning talks were a great way to familiarize oneself with the attendees. It also felt like luxury to be in a room with people who have all had practical experience with digital records, and that we all spoke the same language (SIPs, DIPs, AIPs never had to be defined).

- Courtney Mumma, City of Vancouver Archives

The Digital Archivists organized a two day unconference in May of 2011. The “unconference” is a participant-driven meeting wherein attendees are called on to develop the agenda and activities once they arrive in order to address emerging and cutting-edge topics. The AIMS unconference was a gathering of similarly minded people from the US, Canada, and the UK to bring issues and challenges related to stewarding born digital archives to the table. The Archivists hoped that the unconference format would allow participants to share knowledge, experience, and concerns, while learning new strategies and developing new partnerships to help tackle this enormous challenge we all face.

The 27 participants represented libraries, archives, museums, and digital humanities centers. Despite the differences in our institutions, backgrounds, and training, we learned that we not only shared similar challenges, but also the same hopes for collaboration and innovation. Through an unconference wiki the delegates shared information about their role and institution and proposed topics that they would like to discuss in the event. During the event, notes, slides from lightning talks, and links to useful resources were added to the wiki. The wiki remains publicly available at <https://wiki.duraspace.org/display/AIMS/AIMS+Symposium> as a clearinghouse for the information discussed during the event.

The event resulted in two concrete outcomes. First the delegates agreed that they wanted to continue to work together to help the emerging born-digital stewardship community address shared challenges. The delegates agreed to keep up discussions via the Google Group set up prior to the event and to hold bi-monthly chat/video conference calls to continue discussing the following topics:

- Curriculum Development
- Best Practices and Policies
- Tool Development
- Digital Research Communities

In addition, a specific suggestion was made at the unconference to organize a “Day of Digital Archives” similar to the “Day of Digital Humanities” that’s become an annual event with our DH colleagues. Gretchen Gueguen took responsibility for developing this event, which will take place October 6th, 2011. The project blog is found at <http://dayofdigitalarchives.blogspot.com/>. Thirty-seven participants, both unconference delegates and others, representing archives, libraries, museums, and tool developers from the US, the UK, Australia, and Europe are set to participate by either blogging or tweeting about their activities related to born-digital content management on the 6th.

Comments from UK Event Delegates:

Understand what is happening in digital preservation, meet people doing digital preservation things, look for partners for projects.

- Richard Boulderstone, Director eStrategy, British Library

Excellent balance: shorter presentations and more discussion worked very well. Tool demos would be useful

- Ifor ap Dafydd, Development Officer, National Library of Wales

Reassuring that we’re currently asking the right questions (or at least the same questions as everybody else).

- Owain Roberts, Workflow Analyst, National Library of Wales

Feedback from the event was very positive. Many attendees commented that the opportunity to share experiences with a group of professionals who are also engaged in similar tasks was energizing and would impact their continuing work.

UK AIMS event: Revisiting archival principles from a digital preservation viewpoint, London, June 2011

This one-day event sought to replicate many aspects of the unconference. Organized in collaboration with the Digital Preservation Coalition (DPC), the goal was to facilitate discussion with a group of practitioners to look at three core aspects:

- Collection management
- Arrangement and description
- Discovery and access

A series of brief presentations from invited speakers were followed by open discussion among the delegates about practical issues. These ranged from working with depositors, using and integrating third party tools, born digital archives workflow and other aspects. The event was attended by twenty-three delegates representing eighteen institutions including the British Library, the National Archives of Scotland, University of Cambridge, London School of Economics, The (UK) National Archives, JISC, the National Library of Wales and the Bodleian Library, Oxford.

A wiki was created for the UK event, and the program, slides, and notes from the event are online:

<https://wiki.duraspace.org/display/AIMS/AIMS+UK+event>

While both the US and UK events followed a similar theme there were some key differences. Instead of a pre-selected delegate list by the Digital Archivists as in the US event, the UK event was an open invitation to DPC members. The UK event was promoted as a joint AIMS-DPC event with the theme and agenda being selected by the AIMS team and the DPC undertook most of the administration of the event and promoted it to their members. This meant we had a small but highly experienced audience from a range of institutions. This wealth of practical knowledge and the relatively small size of the group encouraged everybody to share experiences and perspectives.

With the UK digital archivist community already established there was not felt to be a need to generate any direct actions from the day, though comments were sought on the nature and format of the event to see whether it could be repeated, possibly with different emphasis, on an annual basis.

Collecting Repositories & E-Records Workshop, Chicago, August 2011

The AIMS partners hosted a workshop in the run-up to the 2011 SAA (Society of American Archivists) Annual Meeting in August. Forty-five participants from the US and Canada explored the challenges, opportunities and strategies for managing born-digital records in collecting repositories. The workshop was organized around the four

main functions of stewardship in the AIMS framework: collection development, accessioning, arrangement and description, and discovery and access.

In addition to presentations by AIMS Project members, several guest presenters showcased case studies from their hands-on approaches to managing born-digital materials. Seth Shaw, from Duke University discussed the evolution of electronic record accessioning at Duke University and his development of the Duke Data Accessioner. Gabriela Redwine discussed work done in arrangement and description at the Harry Ransom Center at the University of Texas at Austin. Finally, Erin O'Meara showcased work done at the University of the North Carolina at Chapel Hill to facilitate access to born-digital records through finding aid interfaces.

In between presentations, the participants engaged in lively discussions around provocative questions and hypothetical scenarios. At the end of the event, the AIMS partners felt they had gained just as much from the day's activities as they hoped the participants had. The Ideas discussed and case study examples presented played a major role in the development of this white paper:

The program for the event is available on the AIMS wiki:

<https://wiki.duraspace.org/display/AIMS/AIMS+Workshop+-+program>

and the presentations from this event are available via the project blog:

<http://born-digital-archives.blogspot.com/2011/09/aimssaa-part-one-crew-workshop.html>

The Digital Archivists delivered a presentation at on the AIMS project on Saturday morning, providing an overview of the project and the unveiling of the AIMS framework, or the four areas identified as key functions in the stewardship of born-digital materials. There were over 150 SAA conference attendees in the audience, despite competition from Hurricane Irene's effect on travel schedules, an 8 a.m. Saturday timeslot, and simultaneous SAA presentations from colleagues Michelle Light, Dawn Schmitz, and John Novak on delivering born-digital materials online as well as presentations from the archivists for the bands Phish and the Grateful Dead.

The presentations from the event are available at:

<http://born-digital-archives.blogspot.com/2011/09/aimssaa-part-two-saa-session-502.html>

Continuing Community Involvement

As a result of attempts to engage with and garner feedback from the born-digital community, the AIMS partners embarked on continuing projects to collaborate and exchange knowledge with other professionals.

Hull has been approached by a number of other institutions (including the London School of Economics, the John Rylands University Library (The University of Manchester), the Duke of Northumberland Estate, the Wellcome Library, the East Riding Archives Service and the West Yorkshire Archives Service) as a direct result of their involvement in the AIMS project. These contacts resulted in numerous exchange visits and sharing of work-in-progress which has been mutually beneficial to all parties and will continue beyond the life of the project.

Mark A. Matienzo, Digital Archivist at Yale University, and Bradley Daigle, Director of Digital Curation Services at the University of Virginia, will both serve on the development advisory group for the BitCurator project, funded by the Andrew W. Mellon Foundation. BitCurator seeks to develop an open source digital forensics solution for archives. Archivists at Yale University have also begun collaborating more closely internally; staff at Manuscripts and Archives and the Beinecke Rare Book and Manuscript Library have worked together to create workflows and documentation and to share resources to build their capacity and expertise.

3. Day of Digital Archives

The first Day of Digital Archives took place on October 6th, 2011 as a direct outcome of the AIMS project. The event was modeled on the ongoing Day of Digital Humanities project, which encourages participants from around the world working in Digital Humanities to blog, tweet, or otherwise document what they are doing on a specific day each year. The Day of Digital Archives did the same on October 6th, 2011, creating a record of what the field actually looks like as a way to create a deeper understanding with colleagues, researchers, future students, and the world at large.

The idea was first formed at the AIMS Unconference in May as one way to address the lack of awareness some of our colleagues and users have about work with born-digital archives. Digital Archivist Gretchen Gueguen of the University of Virginia took responsibility for setting up and managing the blog and marketing the event. Information about the day was circulated at the Society of American Archivists Annual Meeting as well as on listservs related to the field. Prior to the event more than 50 participants contacted Gretchen to participate and registered with the Day of Digital Archives blog. Numerous other participants joined in on the discussion, particularly through Twitter, on the day itself. These participants were not limited just to those working with born-digital archives, but represented many working with digitized materials, some working within the realm of Digital Humanities, and others involved in software design or archival education. The scope of participants stretched outside the United States to Canada, the United Kingdom and Australia.

While participants were not discouraged from blogging on their own platforms, Gretchen made an effort to cross-post these entries to the centralized Day of Digital Archives blog (<http://dayofdigitalarchives.blogspot.com>) in order to have one clearinghouse for contributions. At the end of the day there were 45 posts on the Day of Digital Archives blog and 8 posts linked to on other blogs. The site received more than 3,000 pageviews on the 6th and continues to be viewed daily, albeit at a lower rate. More than 700 messages were tweeted throughout the day with the #digitalarchivesday hashtag by 365 twitterers.

The topics of posts and tweets covered a broad range of activities from early discussions of the need for particular tools to announcements of completed products. Others used the platform to discuss things like education and training, collaborative initiatives, gaps in tools or shared knowledge, or the activities involved in planning or carrying out projects.

The success of the Day of Digital Archives exceeded initial expectations. The volume of participants and the quality of their submissions were both higher than was anticipated. However, several commenters during the day noted that they were surprised that they had not heard about the effort before that day. This was due to the relatively small amount of effort put into marketing the event. Word-of-mouth was the key tool used as marketing for this initial event, and in some sense this might have added to the excitement surrounding the day's activities on Twitter. However, for future events more formal methods of awareness and participation encouragement should be used.

The only issue that still poses a challenge for Day of Digital Archives is a reliable method for archiving the day's activities. Tweets were backed up to two online service providers: TwapperKeeper and the Archivist, but a more



trustworthy solution should be found. Given the small volume of tweets, creating a simple database of them may be feasible for this year, but may not be feasible in the future. The blogging software chosen was Blogger. While this is a major service of Google and unlikely to go away anytime soon, a long-term solution needs to be found. At the very least obtaining a back-up of the posts and comments should be sufficient.



4. Presentations, Conferences, and Publications

Presentations and Conferences attended on behalf of AIMS

Awre, Chris. "The Hydra Initiative: Underpinning Repository Interaction for Research Support." Paper presented at Fedora UK & I/EU User Group Meeting, Oxford, UK, December 2009.

Wilson, Simon. Attendee at Digital Preservation — The PLANETS Way, London, UK, February 9-11, 2010.

Gushee, Elizabeth. "AIMS." Poster presented at New Horizons in Teaching and Research Conference, Charlottesville, VA, May 5, 2010.

Wilson, Simon. Attendee at European Conference on Digital Archiving, Geneva, CH, April 28-30, 2010.

Edwards, Glynn. "Born-Digital Material @ Stanford" Pecha Kucha session at Northwest Archivists, Inc. Western Roundup, Seattle, WA, April 30, 2010.

Gushee, Elizabeth. "Assessing & Accessing Archival AV Content at the University of Virginia." Presented at the Washington Conservation Guild, May, 2010.

Daigle, Bradley. "AIMS Update." Presented at the Maryland Institute for Technology in the Humanities Computer Forensics and Born-Digital Content in Cultural Heritage Collections Symposium, College Park, MD, May 2010.

Matienzo, Mark. Attendee at the Maryland Institute for Technology in the Humanities Computer Forensics and Born-Digital Content in Cultural Heritage Collections Symposium, College Park, MD, May 2010.

Wilson, Simon. Attendee at Practical Approaches to Electronic Records, Dundee, UK, May 21, 2010.

Wilson, Simon. "Brief introduction to the AIMS Project." Presented at Digital Lives Research Seminar, British Library, London, UK, July 5, 2010.

Wilson, Simon. "Brief introduction to the AIMS Project." Presented at CALM Digital Records Workshop, London, UK, July 22, 2010.

Matienzo, Mark. "Accessioning, Transfer, and Ingest Workflow for Born-Digital Archives in Collecting Repositories." Poster presented at Society of American Archivists Research Forum, Washington, DC, August 2010.

Wilson, Simon and Malcolm Howitt. "Managing Digital Archives: A Calm Perspective." Presented at Society of Archivists Annual Conference, Manchester, UK, September 2, 2010.

Awre, Chris. "Hydra" Presented at Repository Fringe, Edinburgh, UK, September 3, 2010.

Wilson, Simon. "Creating a born-digital workflow that includes both CALM & Fedora." Presented to the National Library of Wales and Archives Wales, Aberystwyth, UK, October 11-12, 2010.

Chan, Peter, Glynn Edwards and Michael Olson. "'Archiving' Digital Lives: Choices, Challenges, and Change." Presented at Digital Library Federation Fall Forum, Palo Alto, CA, November 2, 2010.

Edwards, Glynn, Peter Chan and Michael Olson. "Born-digital Materials." Presented to San Jose State University Library School, San Jose, CA, January 2011.

Matienzo, Mark. "Fiwalk with Me: Building Emergent Pre-Ingest Workflows for Digital Archival Records using Open Source Forensic Software." Presented at Code4lib 2011, Bloomington, IN, February 9, 2011.

Matienzo, Mark and Amelia C. Abreu. "Archival Sensemaking: Personal Digital Archiving as an Iteration." Presented at Personal Digital Archiving 2011, San Francisco, CA, February 24-25, 2011.

Wilson, Simon. Attendee at UK Archives Discovery Forum, London, UK, March 2, 2011.

Burg, Judy. "What will survive of you is...Pencil, paper, pen-drive" Presented at Society of Authors, Northern Region Meeting, Hull, 26th March 2011.

Wilson, Simon. "Born-digital archives & the AIMS Project." Pecha Kucha session at Digital Collaboration Colloquium, Sheffield, UK, March 29, 2011.

Edwards, Glynn and Michael Olson. "Born-Digital Materials in Collecting Repositories: Getting off the Ground." Presented at Society of California Archivists Annual General Meeting, San Jose, CA, April 2011.

Wilson, Simon. "Born-digital archives @ Hull: early steps and lessons learnt (so far)." Presented at Digital Preservation Roadshow, York, UK, April 14, 2011.

Edwards, Glynn, Peter Chan, and Michael Olson. "Born-Digital 'Papers' at Stanford: Overview" Presented at SULAIR Chalk Talk, Palo Alto, CA, May 2011.

Daigle, Bradley, Peter Chan, Gretchen Gueguen, Mark Matienzo, and Simon Wilson. Organizers and attendees at AIMS Unconference, Charlottesville, VA, May 13-14, 2011.

Burg, Judy and Wilson, Simon. Organizers and speakers at Revisiting Archival Principles from a Digital Preservation Viewpoint: joint AIMS / Digital Preservation Coalition event, London, UK, June 10, 2011.

Wilson, Simon. Attendee at Curator's Workbench Workshop, British Library, London, UK, June 30, 2011.

Edwards, Glynn. "Processing Born-Digital 'Papers' at Stanford." Presented at RBMS Pre-Conference, Baton Rouge, LA, June 22, 2011.

Wilson, Simon. Attendee at Preserving Email, London, UK, July 29, 2011.

AIMS Working Group. Organizers and presenters at CREW – Collecting Repositories & E-records Workshop, Chicago, IL, August 23, 2011.

Chan, Peter. "Using Forensic Software to Assign Metadata to Born Digital Archives" Presented at Metadata and Digital Object Roundtable at the Society of American Archivists Annual Meeting, Chicago, IL, August 24, 2011.

Chan, Peter, Gretchen Gueguen, Mark Matienzo, and Simon Wilson. "Born-Digital Archives in Collecting Repositories: Turning Challenges into Byte-Size Opportunities" Presented at Society of American Archivist Annual Meeting, Chicago, IL, August 27, 2011.

Wilson, Simon. "Archivist to Digital Archivist" Presented at Archives and Records Association Annual Conference, Edinburgh, UK, September 1, 2011.

Awre, Chris. "Hydra (and Fedora) in Hull" Presented at Fedora UK & I User Group, Manchester, UK, September 15, 2011.

Publications

Edwards, Glynn. "Born-Digital Material at Stanford," *Archival Elements: Newsletter of the Society of American Archivists Science, Technology, and Health Care Roundtable*, Summer 2010.

Awards

The AIMS project was awarded *Archive Pace Setter* status, part of a program led by the Archives and Records Association (UK & Ireland) in partnership with a number of strategic bodies working across the archives sector. The award recognizes the project's innovative nature and its adherence to good practice in relation to project planning, management, and evaluation.