Introduction

The introduction of desktop computers, MD5 checksums, hand-held devices, and digital forensics into archival repositories brings with it a transformation of accessioning procedures, processing practices, preservation tactics, and research service approaches. The effects of these changes are being felt not only by archivists and librarians but also by researchers and scholars. The arrival of born-digital content into archives has dictated both innovations in archival practice, and promises to bring significant change to research methodologies. As the collections we receive no longer contain just one or two floppy disks, but rather may include complete operating systems and hard drives, archivists must build upon practices developed over recent decades in the handling of electronic records, while reconsidering acquisition procedures and approaches to access.

Using the paper and born-digital materials of the Salman Rushdie Collection housed at Emory University as a case study, this article will explore how the Emory University Libraries, specifically the Manuscript, Archives, and Rare Book Library (MARBL), dealt with these new challenges. The paper discusses each step in the process, from acquisition and processing procedures to the decisions about the type of access we would offer our researchers. We also discuss our approach to integrating both user feedback and studies into MARBL's broader, born-digital archives program.

Soon after acquiring Rushdie's hybrid archive, the library made a commitment to approach the collection as holistically as possible, to prioritize the integration of paper and digital, and to balance the respect for donor concerns with researcher needs. These main tenets informed each decision the team made about handling and processing the digital content. This comprehensive approach to the collection also required that our development of access points and tools embrace both the digital context (i.e., the operating system, original applications, and original file formats) and the larger context of the complete collection (i.e., paper materials and the finding aid). With this goal in mind, the working group developed researcher tools that allow concurrent exploration of emulated environments, item-level, database-driven searches, and the finding aid. Finally, the processes, workflows, and products that comprise Emory's born-digital archives program are contextualized within the human framework of the team itself, which consisted of a collaborative group of technologists, librarians, and archivists.

Background

Emory University acquired the papers of novelist and international figure Salman Rushdie in late 2006. This acquisition was a significant development in Rushdie's relationship with Emory that had begun when he visited the campus in 2004 to deliver the Richard Ellmann Lectures in Modern Literature, a biannual lecture series in which a distinguished writer or critic visits the campus and delivers three lectures and a public reading. Past Ellmann lecturers include Seamus Heaney, whose papers are also at Emory, A.S. Byatt, and David Lodge. In addition to depositing his papers at Emory, Rushdie also began a five-year appointment in the English Department as a Distinguished Writer-in-Residence; he also visits the campus every spring to teach a seminar in the English department as well as deliver several lectures and readings to the University and surrounding community.

Literature is one of several collection strengths of MARBL, and the Salman Rushdie papers joined a cadre that includes poets such as Ted Hughes, W.B. Yeats, and Anthony Hecht, as well as Southern authors such as James Dickey, Flannery O'Connor, and Alice Walker. Rushdie's literary merits are numerous. His second novel, *Midnight's Children*, won the Booker-McConnell Prize for Fiction (now known as the Man Booker Prize for Fiction) when it was published in 1981, in addition to being selected twice as "the Booker of the Bookers," in honour of the prestigious Booker Prize's twenty-fifth and fortieth anniversaries. *The Moor's Last Sigh* (1995), was short-listed for the Booker-McConnell Prize, in addition to winning the Whitbread Novel Award. This book earned Rushdie the distinction of Author of the Year by the British Book Awards.

Rushdie is perhaps most well known, however, for the international attention that followed the publication of his fourth novel, *The Satanic Verses,* in 1988. The book was banned in many Muslim countries for what many believed was its offensive depiction of the Islamic faith and the prophet Mohammed. Iranian religious leader, Ayatollah Ruhollah Khomeini, soon proclaimed that Rushdie and his publishers should be killed. The death sentence – or *fatwa* – sent Rushdie into hiding and was reaffirmed by the Iranian government until 1998.

While the collection consists of over one hundred linear feet of traditional archival material, such as journals, correspondence, and manuscript writings, the reason that this collection stands out from the rest of those housed at MARBL is its large born-digital component. MARBL had received many other collections that included some fugitive computer media,[1] such as floppy disks, CDs, and DVDs, but this was the first time the library acquired entire computers.

It was during the initial negotiations between Rushdie and MARBL that the prospect first arose of including his computers with his papers. With the exception of his very first Macintosh, Rushdie had held on to each of his computers over the years, and he inquired whether or not MARBL would be interested in acquiring the computers as well as the papers. During the *fatwa*, Rushdie had

[1] For a discussion of the term "fugitive media," see Michael Forstrom, "Managing Electronic Records in Manuscript Collections: A Case Study from the Beinecke Rare Book and Manuscript Library," *American Archivist*, vol. 72 (Fall/Winter 2009), pp. 460–77.

become increasingly dependent on his computers and emerging digital technologies that facilitated portability and nearly instantaneous communication, particularly faxing and later, email. In addition, beginning with *The Moor's Last Sigh,* the bulk of his literary output first appeared on the computer screen. With this in mind, MARBL created a proposal outlining its desire to preserve Rushdie's digital files alongside his paper materials. As a result, in late 2006, MARBL received a nearly complete record of Rushdie's digital life, consisting of four computers (one desktop and three laptops), one hard drive (containing files from a fifth laptop that Rushdie had originally planned to give but did not), and several disks that turned out to consist mostly of application files.[2] The choice of acquiring the entire computers and not simply capturing the discrete, user-generated files has enabled MARBL to create innovative access tools that preserve the context in which Rushdie created his literary legacy. The decisions surrounding the way in which we would provide access to Rushdie's born-digital records will be discussed in a later section.

Introduction to the Collection

An overview of the contents of the Rushdie archive is necessary to demonstrate the hybrid nature of this collection. The papers and born-digital materials document Rushdie's professional career, beginning with the publication of his first novel in 1975 through his most recent writings; the materials demonstrate the wide range of his literary endeavours, as novelist, essayist, travel writer, political commentator, defender of free speech, and literary critic. The traditional paper material includes: journals, appointment books, and notebooks; writings by Rushdie, specifically manuscripts and typescripts of his fiction, non-fiction, scripts, and other writings; writings by others about Rushdie in addition to writings by others that concern other subjects; and correspondence, including family correspondence, general correspondence, and correspondence with his literary agents. The materials also include Rushdie's personal papers, such as his passports, photographs from his childhood, and his first prize-winning work (an essay on the Queen's Medal he wrote as a student in 1964). Also featured in the collection are various pieces of memorabilia related to Rushdie, such as buttons, banners, and other objects; and audio and video recordings of interviews, public appearances, and other media events.

The majority of the digital files dates from 1992–2006, and consist of notes and drafts of Rushdie's writings, daily calendars, correspondence, personal and financial files, games, photographs, and downloaded web pages. In interviews

[2] Specifications for this equipment are: Macintosh Performa 5400/180; Macintosh PowerBook 5300c; Macintosh PowerBook G3 [QT9250B5G03]; Macintosh PowerBook G3 [QT9386CEEY8]; SmartDisk FWFL60 FireLite 60GB 2.5" FireWire Portable Hard Drive.

conducted by MARBL staff (to learn more about his digital life), Rushdie stated that he first used his computer as a sophisticated typewriter, but as time passed and technology allowed, he slowly began incorporating all aspects of his life into his computers. This trend became apparent, as later inventories of his computers would reveal that beginning in the mid-1990s, nearly all of the born-digital records overlap with the content and type of material found in the paper portion of the collection. Rushdie explained to Emory Libraries staff that he felt the computer allowed him to more easily organize his literary and personal files. Instead of working at a desk with haphazard piles of papers around him, he could work on a desktop with files that he easily organized into folders, which then automatically sorted themselves alphabetically. He has said that using a computer has made his writing better because it enables him to focus on his writing rather than on the mechanics of writing (typos, page length, etc.).[3]

Forming the Working Group

Shortly after the papers arrived at Emory, the library formed a working group that would be responsible for assessing the new challenges and issues involved in preserving the born-digital material, as well as making it available to researchers in an innovative and responsible way that incorporated both donor concerns and user expectations. This multi-divisional team, the Rushdie Born-Digital Archives Working Group, or BoDAR, for short, included three members from MARBL (Naomi Nelson, Interim Director; Susan Potts McDonald, Head, Arrangement and Description Unit; and Laura Carroll, Manuscript Archivist) as well as three members from the library's Digital Systems Division (Erika Farr, Director, Born-Digital Initiatives; Ben Ranker, Senior Software Engineer; and Peter Hornsby, Software Engineer). This group included a range of expertise, including traditional archival processing, research support, preservation, digital humanities research and methodologies, computer programming, content modelling, and Apple support and programming. With such diversity in skill sets and professional perspectives, it was vital to establish early in the team's work the roles and responsibilities of each member as well as the most effective modes of communication for the team as a whole. It also proved advantageous to include Laura and Susan, who had led the processing of the paper component of the collection, because their familiarity with Rushdie's writing style, works, and life proved essential as we began processing the born-digital component of the collection.

Another important step in this team's formation was developing, and agreeing on, a unified mission and a clear set of desired outcomes. Because of the differences in training and backgrounds, the group first came together with dif-

[3] Salman Rushdie, in discussion with Naomi Nelson and Peter Hornsby, 4 December 2009.

ferent notions of what such a hybrid archive might look like when released in the MARBL reading room. Through conversation, debates, and demonstrations, the group agreed upon the set of driving principles for the program that have been delineated earlier: respecting the hybrid nature of the collection; balancing donor and researcher needs; and providing an authentic research experience. The dialogue that led to these tenets also accomplished the important task of illuminating the unique but complementary skill sets each team member brought to the project.

Surveying the Landscape

As team members began a plan of work for processing, providing access, and preserving the born-digital portion of the collection, the enormity and complexity of the task at hand quickly became apparent. The MARBL Arrangement and Description Unit has developed detailed processing manuals and comprehensive documentation that guide nearly every process from acquisition of the material, to the final delivery of the finding aid on the Internet; policies and procedures for handling electronic records, however, were admittedly less developed. MARBL was not alone. Susan Davis, in her survey of the status of electronic records planning in 125 collecting repositories, found that while nearly 70 percent of respondents had accepted or plan to accept born-digital material, more than 76 percent did not have a policy in place governing the acquisition. Of those reporting the existence of a policy, 57 percent noted that this policy is the same as the one for traditional archival collections. Furthermore, of the fifty repositories answering the question about policies governing preservation and access, 51 percent reported that they had no policy, 30 percent had a policy, and 5 percent stated that their policy was to convert the born-digital records to paper.[4] A review of the literature on managing electronic records in collecting repositories echoes Davis's findings. As she notes in her introduction, the research on electronic records has focused primarily on government institutions and other large institutions or corporations, and often the recommendations rely on the assumption that archivists will be able to have early and frequent interactions with creators to ensure long-term preservation and access to the records.[5] In recent years, there has been a growing movement to address

[4] Susan Davis, "Electronic Records Planning in 'Collecting' Repositories," *American Archivist*, vol. 71, no. 1 (Spring/Summer 2007), pp. 167–87.

[5] See Luciana Duranti, Terry Eastwood, and Heather MacNeil, *Preservation of the Integrity of Electronic Records* (Dordrecht, 2002); Bruce Dearstyne, ed., *Effective Approaches for Managing Electronic Records and Archives,* (Lanham, MD, 2002); Margaret Hedstrom, "Building Record-Keeping Systems: Archivists Are Not Alone on the Wild Frontier," *Archivaria* 44 (Fall 1997) pp. 44–71. See also the following article regarding the Pittsburgh Project: Wendy Duff, "Ensuring the Preservation of Reliable Evidence: A Research Project Funded by the NHPRC," *Archivaria* 42 (Fall 1996), pp. 28–45.

the concerns of other types of archives, those in which the born-digital material often appears without this prior intervention and often exists on fugitive media of questionable provenance. Endeavours such as the Paradigm Project at both Oxford and Manchester universities have sought to research and recommend best practices for institutions that collect private personal and organizational, born-digital material.[6] Several case studies have also addressed how certain collecting institutions have attempted to process, provide access to, and preserve born-digital collections.[7] Many of these and other institutions have contributed to a wider discussion of born-digital records, presenting at conferences, forming informal email discussion groups, blogging, and attending pre-conference gatherings.[8] While both the body of literature and community of practice have grown over the last decade, at the time of MARBL's acquisition, few other archives had acquired hard drives or computers, and the landscape was wide open for new innovations in access and preservation.

Processing the Born-Digital Component of the Salman Rushdie Papers

In the early stages of the acquisition process, MARBL negotiated with Rushdie to establish restrictions on certain portions of his papers, as the collection included a significant amount of personal, financial, and other sensitive information. The existence of these restrictions shaped much of the planning and workflow for this project. Throughout the process, MARBL sought to balance

[6] Susan Thomas, Renhart Gittens, Janette Martin, and Fran Baker, *Workbook on Digital Private Papers, 2005–2007*, Paradigm Project, available at http://www.paradigm.ac.uk/workbook/index.html (accessed 12 February 2011). Other projects include the InterPARES Projects, http://www.interpares.org/; the Digital Lives Project, http://www.bl.uk/digital-lives/index.html; the FutureArch Project, http://www.bodleian.ox.ac.uk/beam/projects/futurearch; and the AIMS (AIMS – Born Digital Collections: An Inter-Institutional Model for Stewardship) Project, http://www2.lib.virginia.edu/aims/ (all accessed 24 February 2011).

[7] Douglas Elford, Nicholas Del Pozo, Snezana Mihajlovic, David Pearson, Gerard Clifton, and Colin Webb, "Media Matters: Developing Processes for Preserving Digital Objects on Physical Carriers at the National Library of Australia," *World Library and Information Congress: 74th IFLA General Conference and Council, 10–14 August 2008*, available at http://www.ifla.org/IV/ifla74/papers/084-Webb-en.pdf (accessed on 12 February 2011); Forstrom; Catherine Stollar Peters, "When Not All Papers are Paper: A Case Study in Digital Archivy," *Provenance*, vol. XXIV (Atlanta, 2006), available at https://ford.ischool.utexas.edu/bitstream/2081/2226/1/023-035.pdf (accessed on 12 February 2011); Catherine Stollar and Thomas Kiehne, "Guarding the Guards: Archiving the Electronic Records of Hypertext Author Michael Joyce," *New Skills for the Digital Era*, Case Study 4, available at http://www.archivists.org/publications/proceedings/NewSkillsForADigitalEra.pdf (accessed on 12 February 2011); Chris Hilton and Dave Thompson, "Collecting Born Digital Archives at the Wellcome Library," *Ariadne* 50 (30 January 2007), available at http://www.ariadne.ac.uk/issue50/hilton-thompson/ (accessed on 23 February 2011).

[8] Stewardship of E-Manuscripts: Advancing a Shared Agenda website, http://ils.unc.edu/callee/emanuscripts-stewardship/index.html (accessed on 12 February 2011); *Practical E-records* blog, http://e-records.chrisprom.com/ (accessed on 23 February 2011).

the need to protect Rushdie's privacy, and the privacy of his family and friends with its mission to make material of scholarly and historical value available to its researchers. Many of the restrictions are routine, such as the closing of his legal and financial files until his death. In addition, papers relating to his family are closed until the death of the specific family member, or seventy years from the date of acquisition, whichever occurs first. The other major restriction involves Rushdie's journals. Beginning in 1974, Rushdie kept detailed journals that include dated notes of both a literary and personal nature, and often have related sketches and comical drawings. They document his creative process and often reveal the development of his writings. Rushdie has stated in numerous interviews that he will soon begin work on an autobiography of his life under the *fatwa*; therefore, all journals written after 1989 are restricted.

Finally, Rushdie initially specified that correspondence from a select number of individuals could be opened only if phone numbers, fax numbers, and home addresses were redacted from the records. Redaction refers to the process of concealing sensitive information and allowing the rest of the information in the record to be viewed by researchers. As processing began, MARBL staff determined that the time, resources, and development needed to effectively redact sensitive information from the correspondence proved too great for the work schedule and resources established for the first phase of processing. Thus, after consultation with Rushdie, MARBL decided to restrict access to all of his correspondence with only a small portion found on his first computer remaining open to researchers. Rushdie also had some specific concerns about the material on his computers. Even in early conversations about researcher access, Rushdie expressed that he did not want the born-digital material to be openly accessible via the Web. In support of these preferences, MARBL decided that the access points we created for Rushdie's digital content would only be available in the MARBL reading room. Finally, the BoDAR team agreed that for ethical and professional reasons we would not attempt to recover deleted files on Rushdie's computers. The team reached this decision after considering a number of factors; the nature of this collection and sensitivity of some of the material, coupled with the concerns that Rushdie expressed about his privacy and the privacy of his family and friends, persuaded the team that data recovery would not be appropriate for Rushdie's digital archive. We will address these issues with donors on a collection-by-collection basis, and make decisions about data recovery for each collection that benefit the donor as well as MARBL and its researchers. A clear understanding of the restrictions within Rushdie's collection was imperative as the working group developed its plan of work. Many of the steps in the process outlined below are there precisely to deal with the rather complex set of restrictions and security issues.

Peter Hornsby, a system engineer in the Digital System Division, conducted the first step of the process involving the initial assessment of the born-digital material and preparation of files for processing. Peter performed triage on the digital collection, which consisted primarily of identifying the physical space to store and appraise the equipment, and inventorying the acquired hardware and storage media. The outcomes of this triage included a detailed assessment of the physical condition of each piece of hardware, such as serial numbers, processing type and speed, and a provisional account of data stored on the equipment. The working group compiled this information in detailed spreadsheets for each piece of hardware and included photographs for documentation.

The next step involved retrieving and duplicating the data. From the earliest stages of the process, the working group relied on a particular content model (see Figure 1). The first silo represents the original hardware and data, which remained untouched except to recover and duplicate the initial data set. The second silo represents what is referred to as the "dark archive." There are only two copies of this dark archive (or master data set), with one stored in a secure, off-site facility more than seventy-two kilometres from MARBL. Only select individuals from the working group have access to this material. Staff never work directly with this data set; it is only used to pull additional copies into the "gray archive." The gray archive represents the working repository of the data, with which the staff can review the material. Finally, the "white archive" represents the fully processed files that are available to researchers, including the redacted files,[9] and files that were migrated from their original format to PDFs for researcher access. The restricted files are not visible in this version of the data.

[9] While BoDAR decided to not pursue redaction for most of the digital collection, the team did elect to redact email addresses in the correspondence included from the Performa 5400. The limited number of email messages and the ease of searching for email addresses allowed the team to pursue this limited redaction.

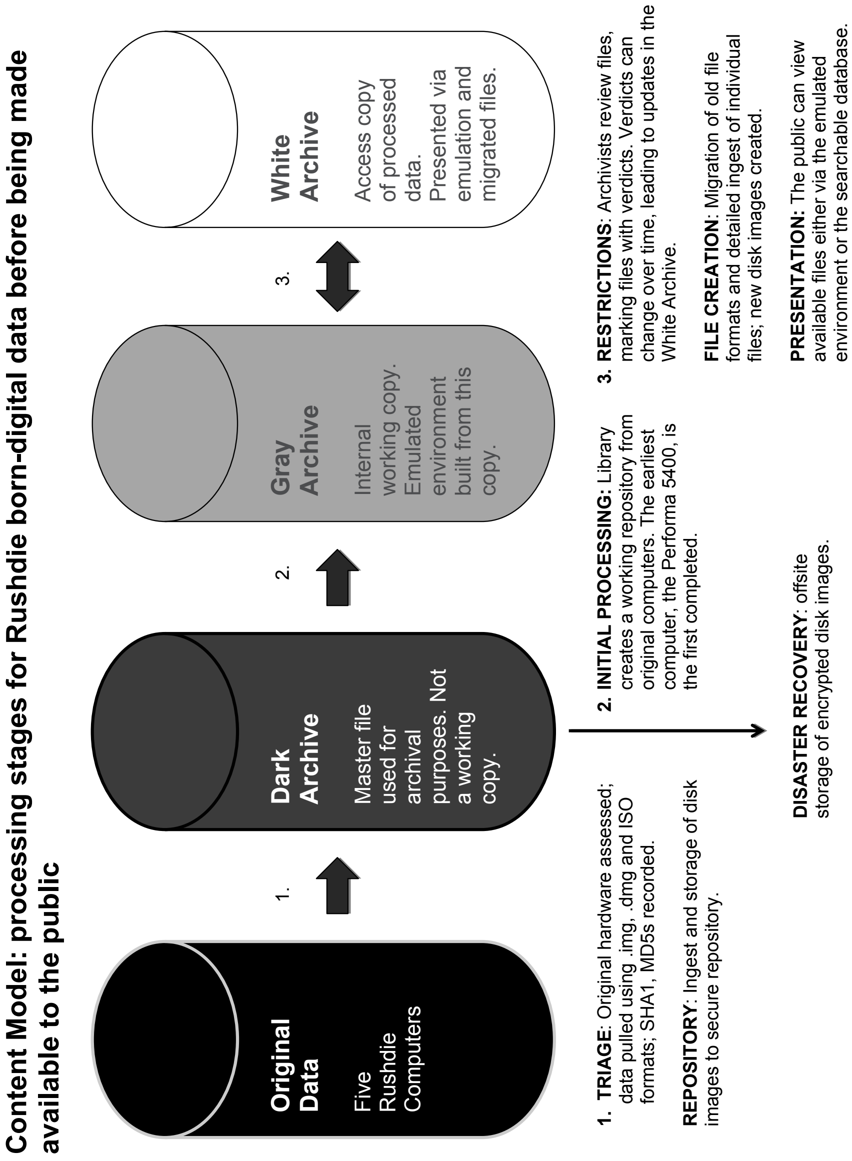
Content Model: processing stages for Rushdie born-digital data before being made available to the public
Original Data
Five Rushdie Computers
1.
Dark Archive
Master file used for archival purposes. Not a working copy.
2.
Gray Archive
Internal working copy. Emulated environment built from this copy.
3.
White Archive
Access copy of processed data. Presented via emulation and migrated files.
1. TRIAGE: Original hardware assessed; data pulled using .img, .dmg and ISO formats; SHA1, MD5s recorded.
REPOSITORY: Ingest and storage of disk images to secure repository.
2. INITIAL PROCESSING: Library creates a working repository from original computers. The earliest computer, the Performa 5400, is the first completed.
DISASTER RECOVERY: offsite storage of encrypted disk images.
3. RESTRICTIONS: Archivists review files, marking files with verdicts. Verdicts can change over time, leading to updates in the White Archive.
FILE CREATION: Migration of old file formats and detailed ingest of individual files; new disk images created.
PRESENTATION: The public can view available files either via the emulated environment or the searchable database.

Figure 1. Content Model Used in Rushdie Workflow

To retrieve the data from the hard drives of each machine, Peter created a disk image of each hard drive. A disk image is an exact replica of a hard drive, bit by bit. The disk image contains all files from the original drives, including user-generated files, applications, and system files. Next, Peter calculated and recorded the checksums for each file.[10] Those working with the data can regenerate the checksum on the data set at any time and compare it against the stored checksum to verify that data has not been corrupted or lost during subsequent migrations or transfers between systems. Calculating the checksum for each file makes it possible to verify the authenticity of born-digital holdings.[11] Once the hard drives were duplicated, Peter began to harvest the metadata for the various types of user-generated content as well as the systems and application files, in order to get a more complete picture of the nature and scope of the born-digital material. We learned very quickly that these computers were rich with data.[12] While the initial plan was to create a database to manage and query the metadata during the processing stage, this approach was later deferred, as the time and money to develop a user-friendly front end were not available at the time; instead, metadata was converted to a spreadsheet format. As a result of the initial data assessment, the working group was able to discern the types of applications that Rushdie used (ClarisWorks, etc.), the way in which he organized and named his files, the amount of user-generated content versus system files for each computer, and other valuable information about the digital content.

Even a cursory review of the data revealed the extent to which Rushdie had backed up his files and had duplicated files across computers. In fact, it was at this point that we learned we had an even more extensive record of his digital life than we thought. While we did not receive Rushdie's first computer, we discovered that he had saved many of the files from that older computer into a folder labeled "OLD MAC," and copied them onto his next computer, the Performa 5400, which was the earliest of the machines we acquired. In anticipation of the review process, we knew that we did not want to review the same file more than once. By using the MD5 checksums as unique identifiers and through the use of Excel formulas, we were able to filter out the duplicate files and create a master list of unique, user-generated files across all of Rushdie's computers.

It was at this point in the process that the working group began grappling with possibilities for providing access to the born-digital material. While there had been speculation about the myriad ways that we would offer this material

[10] Each file has a thirty-two character-long identifier associated with it (generated by a checksum algorithm such as MD5), which can be likened to a unique fingerprint; each file has its very own identifier that cannot be duplicated.

[11] MD5 Message-Digest Algorithm, http://tools.ietf.org/html/rfc1321 (accessed on 28 February 2011).

[12] The computers and hard drive held approximately 11,350 user generated files, consisting of 12,205 MB of data.

istration to make better decisions about processing priorities and approaches. Most importantly, record all of your findings and decisions, and keep these documentation files in an accessible but secure location.

Collection-Specific Processing

Archives should develop a flexible and collection-specific approach to processing and providing access to hybrid and born-digital collections. Archivists can rely on many of the same criteria used for paper collections that have informed decisions about appraisal, processing prioritization, and arrangement and description. These principles still apply, regardless of format. Some collections will warrant item-level description, emulation, and/or a full text searchable database, while others may not. Criteria that can be used to assess the amount and type of arrangement and description, and subsequent access options should be familiar to archivists, including historical or literary value, quantity of material, type and formats of software and hardware, anticipated use, and institutional commitment.

Tool Development

This pressing need underscores the importance of cross-institutional collaboration as the scale of required tool and standards development far surpasses the resources and expertise of any single institution. The early tools that have emerged (e.g., Archivematica[46]) rely on collaborative efforts for development, testing, and implementation, and even those tools are still in their infancy. As more diverse organizations – large or small, public or private, technologically enabled or restricted – begin working with born-digital materials, tools for digital archives will hopefully become more robust and more varied. A solid solution for a large university archive will not likely meet the needs or resources of a small, specialized library.

In closing, our early efforts at Emory to acquire, process, preserve, and provide access to born-digital materials that are a part of larger archival collections have taught us that inquisitiveness, flexibility, and teamwork will be requirements for any digital archives endeavours we undertake in the future. We recommend that organizations of all kinds should, at the very least, begin developing approaches for born-digital archives that address donor relations, collections policies, and long-term storage. While it seems a modest start, basic documentation that details digital material on hand, how and when it was acquired, and its format, at least establishes a baseline for an institution's col-

[46] For more information on Archivematica see http://archivematica.org/wiki/index.php?title=Main_Page (accessed on 8 June 2011).