A Comprehensive Approach to Born-Digital Archives

LAURA CARROLL, ERIKA FARR, PETER HORNSBY, and BEN RANKER

RÉSUMÉ Ce texte explore comment l'arrivée aux archives de contenu créé dès l'origine sur support numérique a engendré des innovations dans la pratique archivistique et promet d'apporter des changements importants aux méthodologies de recherche. Alors que l'étendue d'une collection se déplace de fichiers isolés (« *discrete files* ») vers des systèmes d'exploitation et des collections numériques complexes, les institutions d'archives doivent se fonder sur les pratiques développées au cours des dernières décennies pour manipuler des documents numériques, tout en repensant radicalement la quantité des acquisitions et les approches relatives à l'accès. Ce texte aborde ces changements dans le contexte des manuscrits et des ordinateurs qui font partie des « manuscrits » littéraires personnels de Salman Rushdie, conservés à la Manuscript, Archives, and Rare Books Library (MARBL) de la Emory University. Très tôt dans le développement du projet Rushdie, la bibliothèque s'est engagée à approcher ce matériel de façon aussi holistique que possible, à donner la priorité à l'intégration du papier avec le numérique et à établir un équilibre entre les besoins du donateur et des chercheurs. Ce texte montrera comment la bibliothèque a développé des outils de recherche qui permettent de repérer à la fois le matériel en format papier et le matériel créé en format numérique, et cela grâce à l'émulation et à la recherche au niveau de l'item à partir de bases de données.

ABSTRACT This paper discusses how the arrival of born-digital content into archives has both dictated innovations in archival practice and promises to bring significant change to research methodologies. As a collection's scope moves from discrete files to operating systems and complex digital collections, archives must build upon practices developed over recent decades in the handling of electronic records while also radically reconsidering the extent of acquisition and approaches to access. These changes are discussed within the context of the manuscripts and computers that comprise Salman Rushdie's personal literary "papers," which are housed in Emory University's Manuscript, Archives, and Rare Book Library (MARBL). Early in the development of the Rushdie project, the library made a commitment to approach the material as holistically as possible, to prioritize the integration of paper and digital, and to balance the needs of donors with those of researchers. The paper will outline how the library developed researcher tools that allow concurrent exploration of the paper material and the born-digital material via emulation and item-level, database-driven searches.

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).

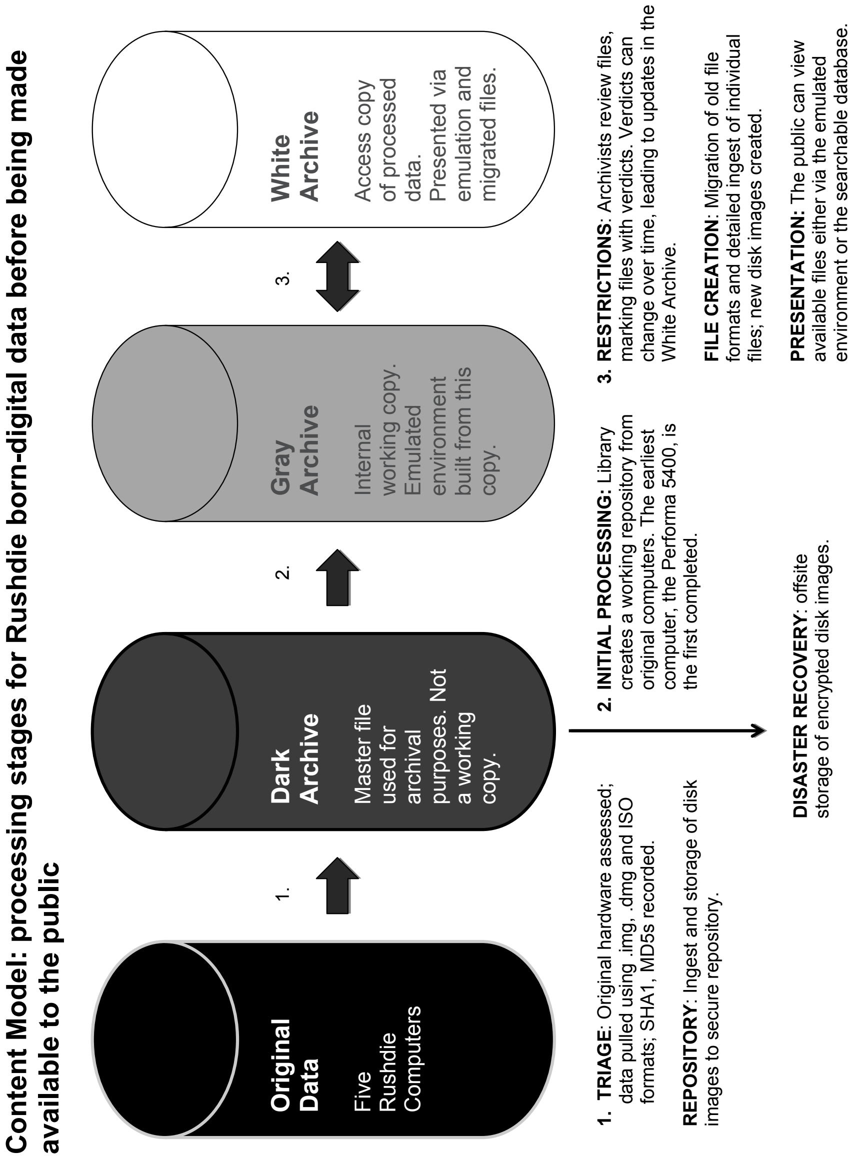
the need to protect Rushdie's privacy, and the privacy of his family and friends with its mission to make material of scholarly and historical value available to its researchers. Many of the restrictions are routine, such as the closing of his legal and financial files until his death. In addition, papers 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.

to researchers in the reading room, the working group made a commitment to preparing two interfaces for researchers to access Rushdie's files and computer environment. The development of these access tools, consisting of both an emulation and a searchable database,[13] occurred in tandem with the processing of the files.

Arrangement and Description Applied to Born-Digital Records

With the final set of unique user-generated files in hand, the manuscript archivist began processing the born-digital material. At the time, we did not know exactly what the access tools would look like or how they would be presented to users, but because of the restrictions involved in the collection, and the sheer number of files involved, a thorough review of the files was necessary.

Numerous questions soon surfaced regarding the way in which we would approach the arrangement and description of the born-digital records. Many archivists and institutions, specifically those featured in the case studies mentioned above, have come up against these same issues. Would we treat the born-digital files the same as print material? Would we impose any sort of arrangement on the materials? How much added description could we provide? As the following discussion will reveal, the application of archival theory and principles may shift to accommodate the differences between paper and born-digital archives, but the underpinnings still guide each step in the process. As Richard Pearce-Moses contends, *what* archivists do will not change, but *how* we do it may in fact change very dramatically.[14]

Soon after Peter had delivered the final set of unique files for each computer in the spring of 2009, we began the process of what can be likened to the "first sort" in manuscript and records processing – grouping materials into broad categories according to type – correspondence, manuscripts, photographs, financial records, etc. In the working copy of the spreadsheet that contained the metadata for all unique files, a new field was created to record this information. The way in which Rushdie organized and named his computer folders and files made this process possible. In fact, Rushdie proved to be quite meticulous in his

[13] When referring to the searchable database we developed as an access tool for Rushdie's digital content, we are using "database" in the sense of a repository of aggregated data that has been indexed for searching and browsing.

[14] Richard Pearce-Moses, "The Perfect and the Possible: Becoming a Digital Archivist," paper presented at the Conference of Inter-Mountain Archivists, 12 May 2006, available at http://www.lib.az.us/about/annualreports/2006/the.perfect.and.the.possible.becoming.a.digital.archivist.pdf (accessed on 12 February 2011); *New Skills for a Digital Era: Proceedings of a Colloquium Sponsored by the National Archives and Records Administration, Society of American Archivists and Arizona State Library, Archives and Public Records*, 31 May–2 June 2006, available at http://www.archivists.org/publications/proceedings/NewSkillsForADigitalEra.pdf (accessed on 12 February 2011).

file-naming conventions and organization. All files associated with a certain work, including notes, drafts, contracts, and publicity, were named and filed with the corresponding title. He rarely used cryptic abbreviations. While not all authors or other content creators will have the same consistent and helpful computing habits that Rushdie exhibited, the ability (and, at times, involuntary necessity)[15] to create his or her own metadata and so easily impose an arrangement scheme prior to transferring material to a repository, highlights important issues and is one tangible advantage of accepting and housing born-digital materials. A key issue this ability raises is the need for pre-acquisition dialogue with donors. Because digital environments allow for customization and personalization, collecting institutions may need to inquire about how the donor managed his or her digital content, approached data transfer between computers, and used applications, directories, and tools. In addition, a donor's ability to assign metadata and arrange content highlights native advantages of digital media, such as how directory structures and file naming can map original order, and demonstrate the donor's own understanding of how one digital object may relate to another (or an entire set of other objects).

As we moved ahead with the processing of the Rushdie files, various criteria were used to assign each file to a corresponding series that mirrored the series we used in the paper material: correspondence, writings, subject files, photographs, and so on. All decisions regarding the rationale for sorting were recorded in a separate sheet in the spreadsheet. As we spent more time with the metadata, we were able to learn about his naming conventions (e.g., his "diaries," which contain files in which he kept a log of his personal daily activities and thoughts). A filter was applied on the filepath/filename field to identify all files that contained these words in the path or file name. Other examples include finding folders with the word "wedding," or "birthday party," and assigning them to the personal papers series; all of his correspondence was filed into folders labeled "letters."

As we expected, the majority of the files on Rushdie's computer relate to his writings, and most of those to his novels. The main words for all of his novels were used to filter these out; other words that were used as folder names and, thus, used as filters include: "column(s)," "non-fiction," "writing(s)," and "work in progress." At this stage, without looking at the files, only so much information could be inferred. For example, the folder entitled "literary matters" could have contained correspondence with agents, proofs, or contracts. Or "wedding" could be the name of a short story. In addition to using file and path names as clues to the type of material, file types were also used (e.g., all files with the

[15] File names, date stamps, and other metadata associated with distinct files are all but unavoidable in a digital environment, and provide useful information to archivists and researchers alike.

extension .jpg were classified as photographs).

This process enabled the archivist to provide the rest of the team with more details about the information on Rushdie's computer. For example, we were able to estimate that out of the approximately 8700 unique user-generated files, 1300 of those were likely manuscripts or files containing some aspect of his writing. This breakdown of the material helped to inform and guide the plan and workflow for the upcoming months. It is important to note that because of the restrictions, the working group decided that each file should receive at least a cursory review. With a more complete picture of the environment, the working group decided to focus on opening the material on the first computer, the Performa 5400, in preparation for the opening of the papers and an exhibit planned for February 2010. It was the earliest of the computers MARBL received, and because it also had the files Rushdie had backed up from his earliest computers, the group decided it was a logical place to begin.

Once the initial assessment of the material was complete, the actual review process began. Through the use of the SheepShaver emulation software (explained in detail below), Laura was able to view each file in its original application software. The working copy of the disk image was refreshed every time she logged on to ensure the data she reviewed and processed was a true and authentic replica of the original hard drive. Each file was evaluated on two levels. First, the contents of each file were compared with the initial assessment of the file type; if the initial estimate was incorrect, then the metadata was corrected and the file was given a new series designation. Secondly, we had to evaluate whether a file contained restricted information or material that would need to be redacted. Each file was assigned a "verdict":

- As is – the file can be released "as is" for both the emulated environment and the searchable database;
- Redacted – the file will need to be redacted for access; it will not be available for emulation, but it will be available in the searchable database;
- Restricted – the file will be restricted and not be accessible in either environment;
- Virtual only – these files will only appear in the emulated environment; they will not be in the searchable database.

A primary example of this type of file are the files that Rushdie may have received from his agent or publisher, entitled "Format," but when opened contain no readable or searchable content.

At this point in the process, both the archivists and the working group as a whole had to make some difficult decisions about the amount of added description we would provide to enhance the user experience. For archivists who regularly process personal papers, these are not new issues and are relevant whether or not the records under consideration are paper or born-digital. The level to which we arrange and describe our material must be in keeping with the resources at our disposal. As our deadline loomed, we had to decide to what extent we would describe these records. Yes, we were dealing with a record creator whose file structure and file-naming conventions made identification and arrangement relatively easy, but how often will this be the case? One of our main goals as archival professionals is to enable our users to find what they are looking for in our repositories. We also believe that with the prospect of hybrid collections in which there is substantial amounts of both paper and born-digital material, users should have the most seamless and coherent experience as possible. We envisioned access tools that would be able to sort, group, and present the data in categories that both experienced archival researchers and those new to the archives would recognize and understand.

In the paper component of the collection, Rushdie's writings are arranged and fully described to the item level.[16] We set our sights on this level of description, knowing that it takes valuable time and money, but with the knowledge that most of our users expect this degree of access. While the extracted metadata from the born-digital files may provide some context (i.e., the file path and file name), it does not resemble the information typically found in the traditional finding aid that a user encounters while doing research. Instead of the full title of an article, short story, or book review, we can provide the date the file was created and the application in which the file was created, along with the folder in which Rushdie placed it. Is this abbreviated description going to be enough for our users? Will they be able to find what they need? Because of time and staffing constraints, the working group decided that we would only be able to provide the next basic level of description for each file, that is, to assign it a subseries designation that would directly correlate to the paper collection arrangement. For example, each file designated as a manuscript would get a further designation, either fiction, non-fiction, script, or other writings.

Describing Born-Digital Content

One of the most exciting aspects of born-digital archives and receiving entire computers and not just files on a disk, is that we can provide numerous points of access. Users can observe and review the files in their original context and order, while also being able to search and sort via various metadata fields that are either automatically harvested or supplied by an archivist. The question still remains: How do we accurately communicate to our users what we have in these hybrid collections? There is some discussion of describing born-digital content in personal papers in the current literature. Michael Forstrom, in his *American Archivist* article, discusses the use of archival description as a model

[16] See http://pid.emory.edu/ark:/25593/8zv36 in the EmoryFindingAids database (accessed on 8 June 2011).

for assessing and maintaining the authenticity of copies of electronic records in manuscript collections. His study focuses on rules in the American standard, *Describing Archives: A Content Standard* (DACS), applied to description in finding aids. As Forstrom notes, there are several elements that are likely to be used in the description of born-digital records, including title, name of creator, acquisitions, and related materials elements. In addition, the Notes field [Section 7.1.4] should include details of any migration or reformatting activities undertaken by the repository since the transfer to archival custody.[17]

There is much less literature, likely because the circumstances are just starting to arise, about how to integrate the description of born-digital records into our current practices – primarily our container lists – to truly represent the hybrid nature of these types of collections. We are currently limited in this project because we cannot link the digital objects to our EAD finding aid due to the collection's restrictions. For now, the scope note includes a section on the born-digital component for each relevant series, with instructions to consult the researcher workstation located in the MARBL reading room for more material. We also considered treating the born-digital records as a distinct format type (as we do with audiovisual material or photographs): creating a "born-digital" series and describing the entirety of the material separately. We abandoned this option, however, and opted to incorporate description of the born-digital material under the nature of the information the file contains, such as his writings, or correspondence.

Accessing Digital Archives

Because Emory Libraries received Rushdie's computers with the acquisition of his personal papers, Emory staff enjoyed the advantages of holding all of the digital content on site, allowing us to fully conceptualize our plans for the digital archive before we began processing. As mentioned earlier in this article, a series of early conversations among the members of BoDAR resulted in several tenets that would inform how we would process and provide access to Rushdie's hybrid collection. The first of these is a commitment to effectively balance the donor's need for data security and personal privacy, with the needs of current and future researchers for deep and effective access to rich primary materials. The second recognizes the importance of representing the paper and digital components as seamlessly as possible, and by providing the fullest possible context for the born-digital content. Finally, Emory Libraries prioritized effective access to the born-digital components of the hybrid archive rather than relegating the needs for researcher access to a secondary or tertiary concern.

These principles informed the workflow and best practices for accessioning

[17] Society of American Archivists, *Describing Archives: A Content Standard* (Chicago, 2004).

and processing the data as well as Emory's research and development efforts to create useful and authentic points of access. Because the majority of the digital content existed within dated systems and unstable hardware, Emory considered two options that address both preservation and access: migration and emulation. Consideration of these two approaches has been admirably undertaken by Erik Oltmans and Nanda Kol, Stewart Granger, and Jeffery Rothenberg, among others.[18] In these studies, distinctions are made between the prioritization of the object in migration and the system in emulation, while diverse arguments are made about the relative feasibility and sustainability of both options.

Within the realm of born-digital archives, migration offers archivists, curators, and others the possibility of identifying an obsolete or at-risk file format and transferring the content of that object to a more stable, current file format. As Oltmans and Kol observe, migration is a preservation effort that "focuses on the digital object itself and aims at changing the object in such a way that software and hardware developments will not affect its availability."[19] Migration has received less emphasis in the literature because proponents of emulation have published more often and more prominently; nonetheless, migration does have its advantages. This preservation tactic is well-suited to collections of "document-like" objects with little dynamic styling or functionality and high consistency; it is vital for collections on degrading media, such as film and audio recordings; and has been put to effective use on corporate and government documents.[20] Clifford Lynch asserts that migration is an approach "that is full of obligations for constant attention."[21] While not a ringing endorsement for migration, Lynch's assessment underscores the vigilance required of this tactic, as there can be intervals of opportunity before a file format becomes completely obsolete that, if missed, renders the digital object illegible.

Born-digital content introduces a debate in the archives between the need to

[18] See Erik Oltmans and Nanda Kol, "A Comparison Between Migration and Emulation in Terms of Costs," *RLG DigiNews*, vol. 9, no. 2, http://www.rlg.org/en/page.php?Page_ID=20571#article0 (accessed on 1 May 2010); Stewart Granger, "Emulation as a Digital Preservation Strategy," *DLib Magazine*, vol. 6, no. 10, http://www.dlib.org.proxy.library.emory.edu/dlib/october00/granger/10granger.html (accessed on 1 May 2010); Koninklijke Bibliotheek, "Emulate," http://www.kb.nl/hrd/dd/dd_projecten/projecten_emulatie-en.html (accessed on 15 May 2010); Gregory Miura, "Emulation Expert Meeting," *International Preservation News* 40 (December 2006), pp. 39–40, http://archive.ifla.org/VI/4/ipn.html (accessed on 1 May 2010); and Jeffery Rothenberg, "Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation" (1999), http://www.clir.org/pubs/abstract/pub77.html (accessed on 11 July 2011).

[19] Oltmans and Kol.

[20] See Catherine Lacken, "Prioritising for Migration of Digital Format in Television Archives," *IASA Journal* 17 (2001), pp. 25–28; and Elizabeth Reuben, "Migrating Records from Proprietary Software to rtf, html, and xml," *Computers in Libraries*, vol. 23, no. 6 (June 2003), pp. 30–33.

[21] Clifford Lynch, "Preserving Digital Documents: Choices, Approaches and Standards," *Law Library Journal*, vol. 96, no. 4 (2004), pp. 609–17.

address pressing preservation challenges and the responsibility to attend to future research needs. The critical analysis of migration as a viable preservation tactic points to the reality that "libraries must consider whether to treat digital materials as artifacts or simply as intellectual content."[22] The debate over content versus context lies at the heart of the challenges of preserving and providing access to born-digital collections. Jeffery Rothenberg argues for the importance of context, and emulation, by articulating how unique digital objects can be: "The meaning of a document may be quite fragile, since meaning is in the eye of the beholder: what may be a trivial transformation to a casual reader may be a disastrous loss to a scholar, historian, or lawyer."[23] Rothenberg's insight into the potential intrinsic value of a digital object articulates the concerns with digital materiality[24] that directs Emory's efforts to present researchers not only with the intellectual content but also with as much of the native digital context of the content as possible.

Just as migration begins with the identification of at-risk file formats and objects, emulation as a strategy must also begin with appraisal and analysis. With emulation, however, the scope of this appraisal is both broader and more technical. Again, it is worth turning to Oltmans and Kol for a succinct description of archival tactics. Emulation, they explain, "focuses not on the object, but on the environment in which the object is rendered."[25] This definition of emulation paired with their account of migration, highlight the issues of authenticity and comprehensiveness that challenge curators of born-digital archives. While migration promises availability across endless shifts in hardware and software specifications, its object-oriented view may be too narrow for archival purposes. The focus on authenticity and context – or paratext[26] – that informs emulation marks it as a promising alternative or complement to migration. Not without its own challenges, emulation involves significant cost, overhead, and development resources, which exemplifies the need for coordinated efforts in the arena of born-digital archives.

Despite the costs involved with emulation, it "promises predictable, cost-effective preservation of original documents, by means of running their original

[22] Andrew Pace, "Coming Full Circle: Digital Preservation: Everything New Is Old Again," *Computers in Libraries*, vol. 20, no. 2 (February 2000), p. 55.

[23] Rothenberg, p. 4.

[24] For a detailed analysis of digital materiality and textual studies see Matthew Kirschenbaum, *Mechanisms: New Media and the Forensic Imagination* (Cambridge, MA, 2008).

[25] Oltmans and Kol.

[26] Paratextual elements "surround [the text] and extend it, precisely in order to present it, in the usual sense of this verb but also in its strongest sense: to make present, to ensure the text's presence in the world, its 'reception' and consumption in the form (nowadays, at least) of the book." Gérard Genette, *Paratexts: Thresholds of Interpretation* (Cambridge, MA, 1997), p. 1.

software under emulation on future computers."[27] The advantages of emulation are centred on its holistic approach. If one emulates an obsolete operating system, the concerns about loss, authenticity, and error that plague migration largely disappear. And, though relatively new to digital preservation, emulation is a well-established practice in various fields, including engineering and computer science, which provides us with both precedent and guidelines.[28] Furthermore, emulation is a practice that is both reversible, in the sense that original data and programming is nearly always stored as backup, and verifiable, because one can test and review an emulation immediately after deploying it.[29]

In addition to the challenge of sustaining a robust emulation program, emulators can also fail to exactly replicate a computing environment or media experience, thus devaluing the emulated environment.[30] As a research tool, emulators also pose problems because it is complicated to have them communicate with other networks and accessory devices, especially current ones, given the fact that the emulated environment is in many ways a technological time capsule. Finally, the question of emulation is not one resolved by a "yes" or "no" verdict. Instead, one must consider at what level to produce the emulation: the file, the application, the operating system, or the hardware.

As the Emory team working on the Rushdie collection explored the advantages and disadvantages offered by migration and emulation, we prioritized our tenet of creating an authentic researcher experience. The BoDAR Working Group believed that Graham Barwell's discussions about originality and authenticity within the fields of textual studies and electronic textuality resonate with born-digital archives.[31] These concerns led to a series of questions that shaped our final decisions on how we would provide access to Rushdie's digital archive. These questions covered such topics as authentic representation, the value of context with born-digital collections, and the importance of digital materiality and paratext. Ultimately, we decided that the context and medium of twentieth- and twenty-first-century archives are of equal importance as those of pre- and post-Gutenberg collections. Scholarly interest in incunabula, early publishing practices, bindings, paper, manuscript hands, marginalia, and front and back matter surely will be mirrored in scholarly research into literary and creative production in the late twentieth century and on. Identifying, categorizing, preserving, and providing access to the materiality of born-digital personal archives can be of equal importance as attending to the content, depending upon the extent, medium, and state of a given collection.

[27] Rothenberg, p. 1.

[28] Lynch, p. 614.

[29] Rothenberg, pp. 19, 28–30.

[30] Lynch, p. 614.

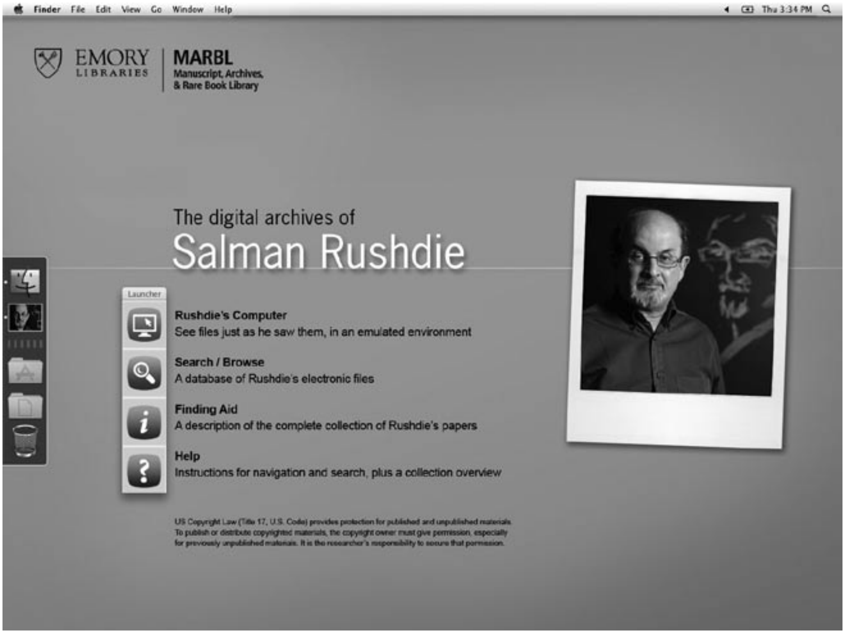
[31] Graham Barwell, "Original, Authentic, Copy: Conceptual Issues in Digital Texts," *Literary and Linguistic Computin*g, vol. 20, no. 4 (1995), pp. 416–19.

Solutions for effectively handling born-digital and hybrid collections are likely going to include migration and emulation, plus additional new practices as they emerge. As is true for most archival endeavours, the chosen tool or technique will depend upon an array of criteria, including institutional capacity and priorities, the merit and nature of the collection itself, and the expectations of the community likely to engage with the collection. Thus, selection, feasibility, sustainability, and co-operation remain key components in debates over migration and emulation, and for the field at large.

Emory's Dual Approach to Access

Numerous factors led Emory's staff to decide on taking a dual approach to access that includes both migration of data and emulation of systems. In particular, the nature of the Rushdie collection demanded a holistic approach, with its inclusion of complete computing environments and its impressive span covering nearly all of Rushdie's digital life. Providing tools built on both migrated data and emulated systems also addressed Emory's commitment to attending to authenticity and usability in its researcher tools and access points. These circumstances and philosophies resulted in our interest in pursuing both tactics at once and prompted us to conceive of a researcher environment that would encompass a range of tools, from the novelty of emulation to the mainstay of archival research, the finding aid.

After the initial organization of the paper records, the Arrangement and Description Unit created a detailed finding aid as a guide to those materials. Some organizing principles reflected in the finding aid informed the next step in development, which addressed the need for comparable tools to access the born-digital content. Emory's solution involved creating a searchable database of PDFs that provides familiar functionality and an emulation of the Performa 5400, which enables the researcher to experience the content and context of Rushdie's digital material in an authentic and contextual way. BoDAR staff strove to incorporate relevant connections between the paper material and digital content, which resulted in the inclusion of the finding aid in the researcher workstation and, more importantly, the importing of organizing principles from the finding aid to the migrated digital content. All three of these access tools – the searchable database, the emulation, and the finding aid – inhabit a single computer workstation in the reading room along with ancillary help documents (See Figure 2).


Finder File Edit View Go Window Help
Thu 3:34 PM
EMORY LIBRARIES
MARBL Manuscript, Archives, & Rare Book Library
Launcher

The digital archives of
Salman Rushdie

Rushdie's Computer
See files just as he saw them, in an emulated environment

Search / Browse
A database of Rushdie's electronic files

Finding Aid
A description of the complete collection of Rushdie's papers

Help
Instructions for navigation and search, plus a collection overview

US Copyright Law (Title 17, U.S. Code) provides protection for published and unpublished materials. To publish or distribute copyrighted materials, the copyright owner must give permission, especially for previously unpublished materials. It is the researcher's responsibility to secure that permission.

Figure 2. Screenshot of Researcher Workstation Available in the MARBL Reading Room.

Most users expect to be able to search by keyword; BoDAR, therefore, decided to develop a searchable database of the "as is" files. Researchers can use it to search both the files themselves and the metadata. The library aspired to create a search tool that would mimic common search tools like online, public access catalogues (OPAC) at academic and public libraries in order to increase ease of use and offer some familiarity. The workflow BoDAR used to create this searchable database involved a file transfer of "as is" files and their metadata from Laura to Ben. Ben then ingested these files and metadata into the Fedora Commons[32] instance we use for the Rushdie digital collection. The next stage of work involved converting these dated Mac files to a more current and accessible file format. Because we wanted the data to be searchable, legible for researchers, and retain as much original formatting as possible, we chose the PDF-A format for the access version of these files. As the conversion process began, it became clear that the workflow would need to accommodate three distinct forms of data: word processing files, fax files, and Eudora email messages. While the process of converting word processing files was fairly simple, con-

[32] For more information about Fedora Commons, see http://fedora-commons.org/ (accessed on 28 February 2011).

verting the fax files and the Eudora email messages proved more complicated. For instance, the faxes were of particular interest because Rushdie used that application regularly but the program was no longer available on the Performa and the files could not be opened with other available applications. Peter had to identify and install a version of the Mac fax application, then export the files as individual TIFFs, and finally compile the individual TIFFs into PDF-A documents that represented the original fax files. The Eudora mail proved similarly challenging and required Ben to produce custom Python software[33] that could read the messages and export them to CERP xml.[34] We opted not to convert this exported CERP into PDF because it was both stable and legible in its initial exported form. Working with outdated and obsolete applications and file formats necessitates both flexibility in approach and creativity in technical problem solving.

Once all the "as is" files were converted to accessible formats and reposited, Ben began connecting this repository data to a search engine that would make it accessible and usable for researchers. The combination of tools we use includes GSearch,[35] which puts Fedora in communication with the search engine, Apache Solr.[36] Ben configured GSearch to read the data and metadata in Fedora and then load it into the Solr search engine, using XSLT.[37] The final stage of this technical development involved Ben writing a Web application using Django[38] that could act as a user interface for the indexed content.

As this Fedora collection of publicly viewable PDF files was being compiled, Emory worked with Resonance, a Web design contractor, to develop a look and feel for the entire researcher workstation, as well as wireframes and graphic design for the user interface we would deploy using a Web browser application on the local machine in the MARBL reading room. Our technical leads, Ben and Peter, then implemented these graphic design components on both the research workstation and the searchable database (see Figures 2 and 3). The deployed version of the searchable database includes keyword searching across the collection of PDFs, browsing by series and sub-series, and brows-

[33] To learn more about the Python programming language, refer to its official website, http://www.python.org/ (accessed on 28 February 2011).

[34] For more information about CERP and CERP xml, see http://www.siarchives.si.edu/cerp/ (accessed on 28 February 2011).

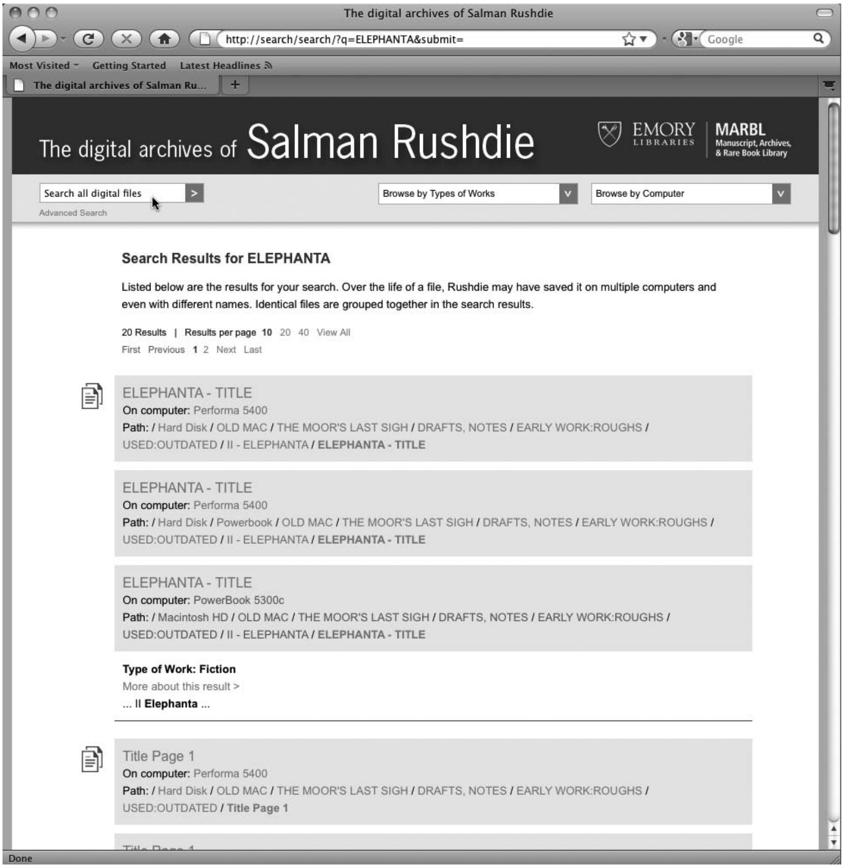
[35] For detailed information about Fedora Commons and GSearch, see http://fedora-commons.org/download/2.2/services/genericsearch/doc/index.html (accessed on 28 February 2011).

[36] More information about Apache Solr may be found at http://lucene.apache.org/solr/#intro (accessed on 28 February 2011).

[37] XSLT is a language that can transform documents encoded in Extensible Markup Language (XML) into other formats or XML documents. For more information, see http://www.w3.org/TR/xslt (accessed on 28 February 2011).

[38] Details and specifications for Django may be found at their official website, http://www.djangoproject.com/ (accessed on 28 February 2011).

ing by file folder and directories. Furthermore, as the remaining computers are processed and made accessible, researchers will be able to browse by computer. BoDAR's interest in keeping the paper components of Rushdie's collection in conversation with the digital files prompted us to include browse categories that echo the organization of the finding aids, while the context-rich emulation tool inspired us to provide directory and folder browsing.


The digital archives of Salman Rushdie
http://search/search/?q=ELEPHANTA&submit=
Google
Most Visited
Getting Started
Latest Headlines
The digital archives of Salman Ru...
The digital archives of Salman Rushdie
EMORY LIBRARIES
MARBL Manuscript, Archives, & Rare Book Library
Search all digital files
Advanced Search
Browse by Types of Works
Browse by Computer
Done

Search Results for ELEPHANTA

Listed below are the results for your search. Over the life of a file, Rushdie may have saved it on multiple computers and even with different names. Identical files are grouped together in the search results.

20 Results | Results per page 10 20 40 View All

First Previous 1 2 Next Last

ELEPHANTA - TITLE
On computer: Performa 5400
Path: / Hard Disk / OLD MAC / THE MOOR'S LAST SIGH / DRAFTS, NOTES / EARLY WORK:ROUGHS / USED:OUTDATED / II - ELEPHANTA / ELEPHANTA - TITLE

ELEPHANTA - TITLE
On computer: Performa 5400
Path: / Hard Disk / Powerbook / OLD MAC / THE MOOR'S LAST SIGH / DRAFTS, NOTES / EARLY WORK:ROUGHS / USED:OUTDATED / II - ELEPHANTA / ELEPHANTA - TITLE

ELEPHANTA - TITLE
On computer: PowerBook 5300c
Path: / Macintosh HD / OLD MAC / THE MOOR'S LAST SIGH / DRAFTS, NOTES / EARLY WORK:ROUGHS / USED:OUTDATED / II - ELEPHANTA / ELEPHANTA - TITLE

Type of Work: Fiction
More about this result >
... II Elephanta ...

Title Page 1
On computer: Performa 5400
Path: / Hard Disk / OLD MAC / THE MOOR'S LAST SIGH / DRAFTS, NOTES / EARLY WORK:ROUGHS / USED:OUTDATED / Title Page 1

Figure 3. A Screenshot of the Searchable Database Available on the Research Workstation in the MARBL Reading Room

For the second important point of access into Rushdie's Performa 5400, the BoDAR staff decided to recreate the original computing environment in which he worked. Emory wants researchers to log in to a digital space that will provide a view into Rushdie's computer exactly as he saw it while researching and writing. Just as researchers of nineteenth-century fiction are interested in book covers and bindings, current and future scholars will be interested in the digital environment that supported Rushdie's literary production. Enabling researchers to securely access processed materials within the original operating system, and to view Rushdie's file directories and desktop required BoDAR to create an emulated environment of the Performa 5400. This research tool acts as an exact replica of the original machine and is populated with files marked by the processing archivist with the verdicts "as is" and "virtual only." This approach provides researchers with an innovative means of exploring Rushdie's digital life.

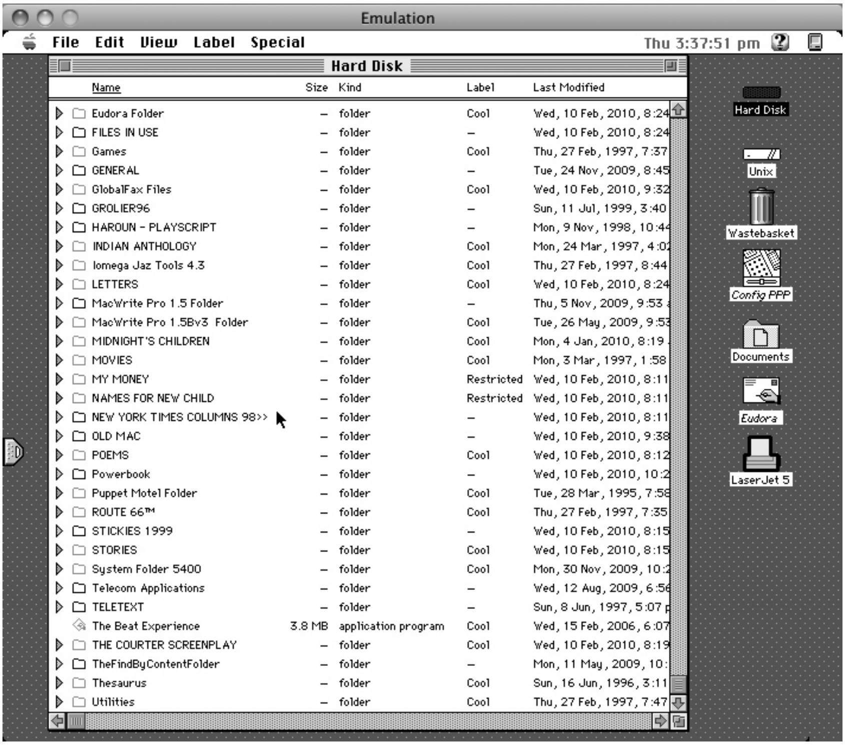
Authentic images of the disk itself fueled our development efforts to create the emulated environment of Rushdie's earliest computer. Peter used SheepShaver, an open source, PowerPC, Apple Macintosh emulator, to emulate the operating system of the Mac Performa 5400. SheepShaver is capable of emulating Mac OS 7.5.2 through 9.0.4 and has been open source since 2002.[39] The software is no longer actively developed by its creators, but they still support it and it is also well supported by a community of hobbyists. The software required some configuring by Peter in order to serve the needs of our project, but his communication with developers and hobbyists provided all the information he needed. Finally, because we own the computer, its operating system, and the ROM chip[40] that is necessary for emulating the disk image, we were not violating any copyright or intellectual property laws.

Once the original environment was effectively emulated, Peter deleted all the user-generated files created by Rushdie, leaving in place only the operating system, software applications, system preferences, and other files and data that would have been pre-loaded on the computer before Rushdie's use. With all of Rushdie's files removed, Peter took the list of verdicts created by Laura and reloaded any files that were marked "as is" or "virtual only" into the emulated environment. To assure authenticity and accuracy, the persistent identifiers associated with each file were used to locate those that were to be transferred back into the emulation. BoDAR decided to take this approach of stripping the emulation and then re-populating it with approved data in order to ensure that no restricted data would remain; this also guaranteed only approved content would make it into the MARBL reading room. With the emulation and its vetted and approved files installed in the researcher workstation, researchers are able to launch an exact replica of Rushdie's Performa 5400 with all of its authentic and at times unstable, mid-1990s Mac attributes (see Figure 4). The error messages that popped up when Rushdie booted up the machine appear on

[39] For more information on SheepShaver, refer to its official website at http://sheepshaver.cebix.net/ (accessed on 28 February 2011).

[40] The role of the ROM chip in emulation and its impact on copyright is further explained in documentation on using SheepShaver and other MAC emulators, which is available at http://www.emaculation.com/doku.php/sheepshaver_mac_os_x_setup (accessed on 28 February 2011).

the screen of the emulator. The applications that failed to launch when Rushdie last used his computer, fail to launch for researchers and archivists. The directory structure, desktop, user preferences, and file naming conventions established by Rushdie are also still intact and await exploration. Though it can be startling to users, the emulation allows researchers to fully interact with Rushdie's digital content: files can be modified, directories can be deleted, and games can be played. It seems that changes are actually made to the data itself, but each time the emulated environment is launched it refreshes the disk image, resetting to the original image. No changes are saved and no modifications are kept.



Name	Size	Kind	Label	Last Modified
Eudora Folder	–	folder	Cool	Wed, 10 Feb, 2010, 8:24
FILES IN USE	–	folder	–	Wed, 10 Feb, 2010, 8:24
Games	–	folder	Cool	Thu, 27 Feb, 1997, 7:37
GENERAL	–	folder	–	Tue, 24 Nov, 2009, 8:45
GlobalFax Files	–	folder	Cool	Wed, 10 Feb, 2010, 9:32
GROLIER96	–	folder	–	Sun, 11 Jul, 1999, 3:40
HAROUN - PLAYSCRIPT	–	folder	–	Mon, 9 Nov, 1998, 10:44
INDIAN ANTHOLOGY	–	folder	Cool	Mon, 24 Mar, 1997, 4:0
Iomega Jaz Tools 4.3	–	folder	Cool	Thu, 27 Feb, 1997, 8:44
LETTERS	–	folder	Cool	Wed, 10 Feb, 2010, 8:24
MacWrite Pro 1.5 Folder	–	folder	–	Thu, 5 Nov, 2009, 9:53
MacWrite Pro 1.5Bv3 Folder	–	folder	Cool	Tue, 26 May, 2009, 9:5
MIDNIGHT'S CHILDREN	–	folder	Cool	Mon, 4 Jan, 2010, 8:19
MOVIES	–	folder	Cool	Mon, 3 Mar, 1997, 1:58
MY MONEY	–	folder	Restricted	Wed, 10 Feb, 2010, 8:11
NAMES FOR NEW CHILD	–	folder	Restricted	Wed, 10 Feb, 2010, 8:11
NEW YORK TIMES COLUMNS 98>>	–	folder	–	Wed, 10 Feb, 2010, 8:11
OLD MAC	–	folder	–	Wed, 10 Feb, 2010, 9:38
POEMS	–	folder	Cool	Wed, 10 Feb, 2010, 8:12
Powerbook	–	folder	–	Wed, 10 Feb, 2010, 10:2
Puppet Motel Folder	–	folder	Cool	Tue, 28 Mar, 1995, 7:58
ROUTE 66™	–	folder	Cool	Thu, 27 Feb, 1997, 7:35
STICKIES 1999	–	folder	–	Wed, 10 Feb, 2010, 8:15
STORIES	–	folder	Cool	Wed, 10 Feb, 2010, 8:15
System Folder 5400	–	folder	Cool	Mon, 30 Nov, 2009, 10:2
Telecom Applications	–	folder	–	Wed, 12 Aug, 2009, 6:56
TELETEXT	–	folder	–	Sun, 8 Jun, 1997, 5:07 p
The Beat Experience	3.8 MB	application program	Cool	Wed, 15 Feb, 2006, 6:07
THE COURTER SCREENPLAY	–	folder	Cool	Wed, 10 Feb, 2010, 8:19
TheFindByContentFolder	–	folder	–	Mon, 11 May, 2009, 10:
Thesaurus	–	folder	Cool	Sun, 16 Jun, 1996, 3:11
Utilities	–	folder	Cool	Thu, 27 Feb, 1997, 7:47

Figure 4. Screenshot of Rushdie's Performa 5400 Emulation Currently Available to Researchers

At the moment for Emory, the ideal approach is to have both migrated data and emulated systems available to researchers simultaneously. These tools can be used in tandem: seek and find in the database, then explore relevant files in their native context through emulation; use the database alone for speedy text searches where digital context is not informative; or engage in digital archaeology by exploring the system itself.

Researcher Needs

Emory thought it especially important to experiment with research access that leverages the digital nature of the material. Borrowing Gérard Genette's print notion of paratext, the Emory Libraries believe that presenting researchers with only the discrete files in migrated form fails to provide the full, digital paratext – the native digital environment and context – of these primary resources. Emulation, on the other hand, empowers the twenty-first-century archival sleuth by providing a much fuller context, offering an insider's view into computing, and enabling research not only into the content but also the technological medium itself and how it might impact literary production.

Emory's commitment to developing authentic and useful researcher interfaces has led to numerous inquiries into information-seeking needs. Before we even began processing the content, we undertook research into studies on information-seeking behaviour and information seeking in archival settings. A model of archival research that has influenced this project is one set forth by Wendy Duff and Catherine Johnson. Duff and Johnson identified four steps in archival research: "1) orienting oneself to archives, finding aids, sources, or a collection; 2) seeking known materials; 3) building contextual knowledge; and 4) identifying relevant material."[41] Absorbing this model and its four steps into the design of the researcher interfaces helped to define needed tools such as interfaces that emphasize context (e.g., the emulated environment), varied help documents, and a database that can be both searched and browsed.

As soon as a prototype of the tools became available, BoDAR staff began a series of internal user tests to assure the functionality and usability of the researcher workstation before its February 2010 launch in the MARBL reading room. This user testing produced findings that prompted both immediate improvements to the interface as well as enhancements that were prioritized but planned for later versions of the tool. For example, of the eight users who tested the researcher workstation, more than half of them were puzzled by empty folders they encountered in the emulated environment. Restrictions on financial, personal, and unpublished material resulted in numerous empty folders on the Performa. Users wondered if the folders were supposed to be empty, or if the system was erroneous or incomplete. In response, BoDAR decided to mark the restricted folders by using colour labels provided in the Mac Operating System (OS), and adding a comment that the folder was restricted in the File Info metadata. This enhancement to the emulation raises some difficult questions. The emulation tool aims to authentically represent the digital environment in which

[41] Wendy Duff and Catherine Johnson, "Accidentally Found on Purpose: Information-Seeking Behavior of Historians in Archives," *Library Quarterly*, vol. 72, no. 4 (October 2002), p. 472.

Rushdie wrote, but researchers' need for clarity in that environment required us as archivists to introduce new data into that emulation. Ultimately, we decided that the researcher's confusion over empty folders justified this intrusion. The colour-coding and noting of restricted content is a direct response to the need for privacy that prompted the removal of the data; BoDAR ultimately decided that one archival intervention (removal of sensitive data) warranted a balancing one (acknowledgement of the missing content).

Current efforts at gathering and responding to researcher feedback involve tracking data about users interacting with the Rushdie materials and then compiling it so that we can compare it with the use of our other collections. The Rushdie collection was one of MARBL's top ten most used collections for 2010 and the numbers promise to be the same in 2011. Of these users, more than half make use of the digital content. Through follow-up interviews and further data tracking we intend to better understand why the digital material is used (or not) and how researchers interact with digital content.

Of the researchers who used the digital material, several engaged in only cursory exploration, wanting to get a sense of the content and the interfaces. Early feedback from some researchers who have used the digital archive more extensively, reveals that they encounter initial difficulty with the emulation mainly because it is an old operating system. We quickly forget the computing assumptions we take for granted. One researcher described his first forays into the emulated environment as "liminal shock," because the overall experience is rooted in a current Mac workstation but his interaction with the content itself is embedded in a mid-1990s OS and some cranky mid-1990s software.

Such responses and reactions raise issues around archival methodologies. The introduction of born-digital content, legacy formats, and obsolete software applications places new demands on researchers. Researchers interested in primary sources from the latter part of the twentieth century and the whole of the twenty-first century, may very well have to modify existing research methodologies or develop new ones, such as critical analysis of emulated environments. Furthermore, tools for researching born-digital content will prompt researchers to develop means of assessing the integrity and authenticity of the archival content these tools deliver. MARBL plans to continue tracking researcher use, conducting follow-up interviews when appropriate, and undertaking user studies to better understand and address researcher needs.

Ongoing Work at Emory

With the first set of data now available to researchers in the MARBL reading room, MARBL staff are anxious to add new functionality and new content to the researcher workstation. Based on continued user studies, the staff will plan for, and implement, new user tools within the researcher workstation (e.g., text analysis tools and data visualization applications) as well as personalized features within the searchable database (e.g., saved searches and personal researcher "bookshelves"). In addition to these researcher tools, MARBL plans to work with various partners to develop efficient and sustainable methods for redacting content so that files containing restricted elements can be redacted and made available to researchers. MARBL also wants to continue processing Rushdie's born-digital content so that within the next few years, the approved content from all four machines and the external hard drive is available for researcher access.

More broadly, the library is currently establishing a digital analysis lab for born-digital archives, which will allow MARBL to more efficiently create disk images, review born-digital content for damaged or compromised dates, filter for file duplication, create and assign MD5 checksums, and review the substantive quality of a born-digital collection.[42] We are also currently shifting the repository infrastructure for digital archives from an isolated repository to one that is included within the larger data management system supported by Emory Libraries, though this shift requires the addition of numerous security measures including encryption and a customized firewall. Because so few archives currently provide access to born-digital or hybrid collections and none other than MARBL currently provide access through emulation, the library is also keenly interested in pursuing user studies into the information-seeking behaviour of the researchers exploring these collections. While novel and unique now, these materials will soon become the mainstay of contemporary archival collections; libraries and archives must, therefore, explore how researchers wish to interact, save, and transport important primary resources of the twentieth and twenty-first centuries.

Conclusions

Throughout the early work on Rushdie's hybrid collection and the acquisition of other hybrid and born-digital collections, Emory's staff have learned some important approaches to acquiring, preserving, and providing access to hybrid and born-digital collections. While many of these findings have been mentioned and discussed throughout this paper, several are worth delineating here: the necessity of collaboration; the need to engage with other fields and communities; the role of pre-acquisition consultations with donors and content creators; the importance of triage and appraisal; the value of collection-

[42] Other groups and institutions are already providing important examples and best practices for such archival forensic workstations. Significant examples include the AIMS grant already mentioned, the work undertaken by Jeremy Leighton John and his colleagues at the British Library, and the recent CLIR report, *Digital Forensics and Born-Digital Content in Cultural Heritage Collections* (2010, Washington, DC) by Matthew Kirschenbaum, Richard Ovenden, and Gabriela Redwine.

specific processes and workflows; and the need for co-operative tool development.

Collaboration

Collaboration both within a given institution and with the larger community is essential to the success of developing sustainable, digital archive programs. Collaborating across divisions at Emory proved vital to the early success of our digital archives efforts, and we recommend any institution undertaking digital archives work to solicit experts in archives, information technology, and academic research for their working group or team. Staffing a hybrid working group is an important form of collaboration, but institutions must also engage with the broader community for support and guidance on developing best practices, standards, and identifying useful tools. Emory's participation in collaborative grants, symposia, and digital archives meetings has provided us with invaluable insights into forensics, workflows, donor negotiations, and researcher expectations.[43] Ultimately, broad and deep collaboration throughout the archival profession will be essential to the success of archives and libraries as they manage, preserve, and provide access to born-digital collections. Software registries, hardware repositories, analysis and processing tools, secure redundant storage, file format registries, and preservation guidelines are but a sampling of the necessary components to archiving born-digital materials. It is neither sustainable nor sensible for any institution to attempt mastery in each and every area.

Learning from Diverse Communities

As new technologies and personal computing habits impact the format of personal and organizational archives, curators and archivists will need to be open to archival solutions and approaches from unexpected fields and communities, such as early computing hobbyists and enthusiasts, among others. For example, standard practices for computer forensics among law enforcement agencies and legal experts have informed early work in acquiring and analyzing disk images for archival storage, processing, and access. Documentation and guidance from Mac enthusiasts advanced our work with Rushdie's early Macintosh computers.

[43] For example, Emory staff participated in an NEH Office of Digital Humanites Start-up Grant, which resulted in important new partnerships and the following paper: Matthew Kirschenbaum (Project Director), Erika Farr, Kari M. Kraus, Naomi L. Nelson, Catherine Stollar Peters, Gabriela Redwine, and Doug Reside, "Approaches to Managing and Collecting Born-Digital Literary Papers for Scholarly Use," a National Endowment for the Humanities Office of Digital Humanities white paper, available at http://www.neh.gov/ODH/Default.aspx?tabid=111&id=37 (accessed on 28 February 2011).

Early digital archives initiatives demonstrate both the need for, and the benefit of, importing tools and practices from disparate fields and disciplines.

Working with Donors

Because of the great potential for content creators to shape their own digital archives, archivists and curators must work with donors before and during the acquisition process. There are several groups and projects that have developed useful tools to work with donors as archival repositories start acquiring more born-digital material. The Andrew W. Mellon Foundation-funded AIMS project – Born-Digital Collections: An Inter-Institutional Model for Stewardship – developed a survey to be used as a prompt sheet for phone or face-to-face interviews with donors by curators or digital archivists.[44] It includes questions that attempt to assess an individual or organization's digital life, and provides a practical template for conversations with donors that can be undertaken by institutions both large and small.[45] Institutions should develop appraisal techniques appropriate to born-digital content that allow the pre-acquisition assessment of digital material and hardware. For example, an institution may decide that it cannot actively collect hardware if there is little or no data housed on it. If a donor reveals that he or she kept the majority of his data on floppy disk drives or an external hard drive, and there are no plans to provide emulation as an access tool, the institution may choose not to acquire the monitor, keyboard, or even the CPU. In addition, the staff member working with the donor should have an accurate sense of what the institution is interested in preserving and capable of acquiring.

Triage and Appraisal

It is vital to invest significant time and effort into the initial stages of appraisal, accessioning, and triage, as the outcomes from these early stages of work can dictate the quality of the processed and accessible collections. For example, the decision not to make a disk image of a hard drive will greatly inhibit your ability to provide emulation as an access option in the future. Depending on your institution's budget and priorities, the initial information gained while accessioning the born-digital material may be the only information available for quite a while. Being able to glean as much information as possible (e.g., file types and the number of files) in that first pass will enable you and your admin-

[44] Another example of the importance of collaboration within digital archives, this project team includes staff from the University of Virginia, Stanford University, the University of Hull, and Yale University.

[45] Available at https://docs.google.com/document/d/1-zhAUIAOyvBmGvmi-jHeQZOLbsObNxt5j8SOZPQAYEo/edit?hl=en_US&authkey=CKnE4ogP (accessed on 8 June 2011).

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).

lections and maps the digital collections themselves. Basic policy development and documentation of digital material, when combined with engagement with the emerging communities around born-digital and hybrid archives, will adequately position institutions to develop their own digital archives program when resources and priorities allow.