Born Digital: The 21st Century Archive in Practice and Theory

Gabriela Redwine
gredwine@mail.utexas.edu
Harry Ransom Center, The University of Texas at Austin

Matthew Kirschenbaum
mkirschenbaum@gmail.com
University of Maryland

Michael Olson
mgolson@stanford.edu
Stanford University Libraries / Academic Information Resources Stanford University

Erika Farr
elfarr@emory.edu
Robert W. Woodruff Library, Emory University

As more people rely on computer technologies to conduct their personal and professional lives, born-digital materials such as emails, Word manuscripts with tracked changes, blog entries, text messages, and tweets will constitute the archives of the future. Archival repositories at places like Stanford University, Emory University, and The University of Texas at Austin have been receiving born-digital materials for over 20 years but have only recently begun working actively to preserve these items in their original digital formats.

As part of this work, archivists have begun to look to other fields, such as computer forensics and law enforcement, for equipment and methodologies to use in the acquisition and preservation of born-digital materials. The application of forensics technology to born-digital content in archives and the development of tools to facilitate access to these materials hold great promise for humanities scholarship and teaching.

This session brings together digital archivists, librarians, and curators to discuss some of the forensic techniques and equipment being used to preserve born-digital archival materials at the Stanford University Libraries, the researcher interfaces Emory University has developed to provide access to Salman Rushdie's computers, and the broader implications of these developments for the concept of "archives" in a variety of disciplines, including information science, literary studies, history, and cultural studies.

Michael Olson, Digital Collections Project Manager for Stanford University Libraries, will begin the session with a discussion of the applicability of forensics software to the acquisition and description of born-digital archival materials at Stanford. Erika Farr, Director of Born-Digital Initiatives at Emory's Woodruff Library, will discuss the researcher interfaces developed for use with Salman Rushdie's computers and the results of user studies currently underway to explore the potential effects of analog-digital hybrid materials on research methodologies and scholarly communication. Gabriela Redwine, Archivist and Electronic Records/Metadata Specialist at the Harry Ransom Center, will consider the computer as an archival object that challenges both archival and scholarly notions of what an archives is and can be, as well as the functions it may serve.

The panel will be chaired by Gabriela Redwine, of the Harry Ransom Center, The University of Texas at Austin. Matthew Kirschenbaum, Associate Director of the Maryland Institute for Technology in the Humanities (MITH), will serve as respondent.

Computer Forensics in the Archive: An Analysis of Software Tools for Born Digital Collections

Michael Olson
mgolson@stanford.edu
Stanford University Libraries / Academic Information Resources Stanford University

Stanford University Libraries hold an increasing amount of digital archival material. This principally comprises magnetic and optical disks and tapes containing digital files produced both via historical computing platforms on legacy media, as well as via contemporary applications on modern media. Analysis of recent acquisitions from the last five years has shown a five-fold increase in the number of collections containing digital archival materials. Without near-term action, these materials are at the greatest risk of loss and are likely to disappear from the corpus of primary source materials. The imminent loss of digital archival materials now confronts curators, digital archivists, and researchers who desire to use and preserve these digital records.

Computing forensics is a discipline that is still very much dominated by the law enforcement community and the need for digital evidence that can be verified in a court of law. It is based on the following core principles: "that evidence should not be altered, examination results should be accurate, and that examination results are verifiable and repeatable" (Pollitt, 1995). These same principles translate to the archival world, where provenance or verifiable custody is a foundation of archival theory. Curators, digital archivists and researchers have the same requirement that documents, whether in an analogue or digital format, be verifiable.

Digital investigations, both criminal and commercial, have driven the development of forensic software tools and training. Commercially produced forensic software and training certification programs are almost universally adopted by law enforcement agencies. Open-source software for the capture and analysis of digital archival materials is available as an alternative, but there is even less data on how these tools work or could be used in the archival field.

Beginning in early 2009, staff from Stanford's Digital Libraries Systems and Services group met with our archivists to assess the preservation and access needs for digital archival materials. Out of these discussions at-risk collections were identified and it was determined that our highest priority was to safely migrate these collections off at-risk media in a forensically sound manner. Our greatest concern was that the floppy disks, magnetic tapes, and hard drives in our collections would degrade before we could develop a comprehensive program to both preserve and make these materials available to researchers. A second priority was to acquire software tools that would allow our archivists to assess the contents of digital materials and develop methods for making them available.

Alongside this priority-setting exercise, Stanford sought advice from the participants at the British Library's Digital Lives Conference. Jeremy Leighton John at the British Library and staff from the Paradigm Project (co-directed by Oxford and Manchester) were particularly helpful in providing their expertise and a list of potentially useful hardware and software (Paradigm, 2005-7). Following up on this advice Stanford began an intensive discussion with multiple forensic vendors that currently supply and train many law enforcement agencies in the United States. These discussions were notable by the surprise many forensic firms expressed when presented with our archival needs; law enforcement is clearly driving the market for forensic hardware and software.

In the summer of 2009, Stanford University Libraries acquired a suite of forensic hardware and software and has undertaken an extensive program to test a wide range of commercial and open-source forensic software applications and evaluate which applications are most appropriate for use by our curatorial staff, digital archivist, and donors. This paper summarizes our experience in evaluating our academic archiving needs against the range of commercial and open-source forensic software applications. It is important to note that our findings are not scientific product evaluations. The results provided in this paper merely reflect our own experience using these different methodologies to forensically image and analyze digital archival materials from the perspective of a curator, a digital archivist, and a potential donor of digital archival materials.

The results of our findings are based on the following criteria: the nature of the archival collection, skills required to use the software effectively, an evaluation of feature sets, potential for integration with existing archival software such as the Archivists' Toolkit, support for metadata outputs and preservation services, application cost, and supported forensic disk image formats. Our non-scientific evaluation includes two of the largest commercial applications used by the forensic law enforcement community: Guidance

Software's EnCase Forensic ™ and AccessData's FTK (Forensic Toolkit) 3.0 ™. In addition, we will include our evaluation of open source software such as The Sleuth Kit and a small number of freely available forensic utilities.

References

Paradigm project (2005-7). *A Proposal for Intellectual Access to Hybrid Archives. Workbook on Digital Private Papers.* http://www.paradigm.ac.uk/workbook/cataloguing/intellectual-access.html (accessed 12 November 2009).

Pollitt, M. M. (1995). 'Principles, Practices, and Procedures: An Approach to Standards in Computer Forensics'. *Second International Conference on Computer Evidence.* Baltimore, Maryland, 10-15 April 1995.

Finding Aids and File Directories: Researching a 21st Century Archive

Erika Farr
elfarr@emory.edu
Robert W. Woodruff Library, Emory University

The introduction of desktop computers, MD5 checksums, handheld devices, and digital forensics into archives and special collections brings with it a transformation of accessioning procedures, processing practices, preservation tactics, and research service approaches. The impacts of these shifts and transformations will be felt not only by archivists and librarians but also by researchers and scholars.

In this paper, I will discuss how the arrival of born-digital content into archives has insisted on innovations in archival practice and promises to bring significant change to research methodologies. As a practical, concrete means of framing this discussion, I will focus on a particular case study: Salman Rushdie's hybrid "papers," housed in Emory University's Manuscript, Archives, and Rare Book Library (MARBL). By considering the acquisition, processing, and accessibility of this collection, this paper will discuss the new challenges introduced to archival science by such hybrid collections. More importantly for the purposes of this paper, user testing and user studies currently underway on the Rushdie materials will provide valuable data and insight into how hybrid collections of primary materials may influence archival research habits and scholarly communication.

The 2006 acquisition of Salman Rushdie's papers, which included both traditional manuscript materials and a series of personal computers, provided Emory University Libraries with its first significant hybrid collection of personal papers. With the exception of a few articles (e.g. Thomas and Martin, 2006) and the *Workbook on Digital Private Papers* produced by the Paradigm project (2005-7), very little documentation existed to guide the staff at MARBL and in Emory's Woodruff Library in its approach to accessioning and handling these materials. Early in the development of the Rushdie project and Emory's Born-Digital Archives program, the team made a commitment to approach the material as holistically as possible, prioritize the integration of paper and digital, and balance donor requests with researcher needs. Such a philosophy prompted us to begin processing by first capturing complete disk images of all five hard disks, then creating verifiable MD5 checksums, and revisiting security and confidentiality concerns at virtually every processing turn. Our comprehensive interest in the collection demands that our development of access points and tools embrace both the digital context (e.g. the operating system, original applications, original file formats) and the larger context of the complete collection (e.g. paper materials and finding aids). This interest in context led us to explore virtualized environments as a point of access and resulted in the development of researcher tools that allow concurrent exploration of emulated environments, the finding aid, and item-level, database-driven searches.[1]

In addition to providing a greater level of detail about the early processing of and planning for the Rushdie papers, this paper will also highlight important early collaborations. In particular, I will discuss some of the valuable insights gained while participating in an NEH Office of Digital Humanities Start-up Grant with partners from the Maryland Institute for Technology in the Humanities and from the Harry Ransom Center at the University of Texas at Austin (see Kirschenbaum et al., 2009). This start-up grant has had important influences on our program development.

As this planning grant was concluding, Emory began finalizing plans for the public release of Rushdie's archive. In preparation for this major milestone in February 2010, staff at MARBL and in the Digital Systems division of Emory's Woodruff Library worked diligently to process the materials in both traditional and more innovative ways and to create tools, infrastructure, and interfaces that will enable effective researcher access to a selection of the born-digital materials as well as the finding aid for the paper materials. With a completed prototype of the researcher workstation ready for initial testing in early October 2009, staff undertook a cornerstone piece of work for the Born-Digital Archives program: user testing. Because born-digital archival content changes how researchers access and interact with materials, it will necessarily result in changes in how researchers undertake their work. In order to provide optimal support of and service for such scholarly pursuits, archives and libraries must relentlessly explore, study, and analyze researcher needs and habits. Given the current transformative period in archives, it is especially important that we know, even anticipate, what researchers will want to do with these materials ten, twenty, even fifty years from now.

In an essay discussing researcher habits in archives, Duff and Johnson argue that archives need more accurate and diverse scenarios of use in order to better understand how scholars use and interact with archival material (2002, p.473).[2] This observed need for more data and better understanding about how researchers currently use archival materials fuels Emory's interest in gathering user feedback on born-digital materials and exploring effective interfaces for such collections. With this mission at the program's core, we will continue to undertake testing and user studies, beginning in earnest in March 2010. Based on findings and results from these studies, we will take the practical steps of revising and augmenting our systems and services, as well as undertaking the slightly more theoretical activity of documenting these habits in order to begin articulating shifting methodologies in scholarly research.

This paper will elaborate on the activity and development of Emory's Born-Digital Archives program, expound on the work involved in providing researcher access to Rushdie's hybrid collection, and introduce early findings from user studies and testing on the initial set of tools produced for the release of Rushdie's hybrid archive. Discussion of these activities within the framework of how hybrid collections impact research and supported by data gathered during studies and testing should begin to illuminate some of the ways in which research may evolve and transform in the twenty-first century archive.

References

Duff, W. M., Johnson, C. A. (2002). 'Accidentally Found on Purpose: Information-Seeking Behavior of Historians in Archives'. *Library Quarterly*. **72(4)**: 472.

Kirschenbaum, M. (2007). 'Hamlet.doc?: Literature in a Digital Age'. *The Chronicle of Higher Education*. **53(50)**: B8-9. http://chronicle.com/free/v53/i50/50b00801.htm (accessed 20 October 2009).

Kirschenbaum, M. et al (2009). *Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use*. NEH Office of Digital Humanities. http://www.neh.gov/ODH/Default.aspx?tabid=111&id=37 (accessed 2 November 2009).

Paradigm project (2005-7). 'A Proposal for Intellectual Access to Hybrid Archives'. *Workbook on Digital Private Papers*. http://www.paradigm.ac.uk/workbook/cataloguing/intellectual-access.html (accessed 30 October 2009).

Thomas, S., Martin, J. (2006). 'Using the Papers of Contemporary British Politicians as a Testbed for the Preservation of Digital Personal Archives'. *Journal of the Society of Archivists*. **27(1)**: 29-56. doi:10.1080/00039810600691254.

Notes

1. In his *Chronicle of Higher Education* article "Hamlet.doc?: Literature in a Digital Age," Matthew Kirschenbaum's description of the rich potential of born-digital papers demonstrates one example of scholarly interest in born-digital material beyond discreet files.
2. Duff and Johnson focus on historians in this piece, but their conclusions and observations pertain to humanities research more broadly.

Archives and 'the Archive': The Computer as Archival Object

Gabriela Redwine
gredwine@mail.utexas.edu
Harry Ransom Center, The University of Texas at Austin

In 2009, the National Endowment for the Humanities funded a project entitled "Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use," which supported site visits among personnel working with the born-digital components of three significant collections of literary materials: the Salman Rushdie Papers at Emory University, the Michael Joyce Papers at the Harry Ransom Center, and the Deena Larsen Collection at the Maryland Institute for Technology in the Humanities (MITH). In the two publications emerging from that project, the grant collaborators, and Matthew Kirschenbaum in particular, articulated the idea of an author's computer as what Kirschenbaum termed a "complete material and creative environment"—one that an author inhabits like she would a suit of clothes, or an office, or a self (Kirschenbaum et al., 2009a and b). This paper will build on that understanding of the relationship between computer and author to consider the ways in which the computer, as a complex archival object, pushes the boundaries of traditional archival practice and also has the potential to reshape the discussion of "the archive" as a subject of critical inquiry.

The forensic techniques Stanford, Emory, the Ransom Center, and other repositories are using to capture images of disks and hard drives offer the potential for archivists to preserve and analyze more information about authors' work and lives than ever before. For example, an author's browsing history could provide insight into her online research during a particular period of creativity, or the trash folder of an email account could contain discarded emails important to an understanding of a particular manuscript. The tools being developed as part of forensic projects like Simson Garfinkel's Real Data Corpus can be used to recover data and characterize relationships between data sets. For example, it would be possible to map social networks between computers and create a visualization showing which authors were communicating with each other during a certain period of time. This type of work could be done using hard drives residing at a single repository or as part of a collaborative project across institutions. But does the existence of these types of materials and the technological capability to preserve and analyze them mean that archivists should? What is potentially hidden or revealed when a laptop or a server-based Twitter or email account is part of a collection acquired by an archival repository? What ethical concerns arise around born-digital manuscript drafts deleted by a creator, files "hidden" within a computing system, or correspondence that exists only in the cloud?

The Society of American Archivists, North America's oldest professional association for archivists, defines an archives as a body of "materials created or received by a person, family, or organization, public or private, in the conduct of their affairs and preserved because of their enduring value" (2005). Questions of value have long been at the center of debates among archivists, scholars, activists, historians, politicians, governments, and others about what gets saved, by whom, and to what end. The implications of historical definitions of archives and the presumably objective role of the archivist continue to inform scholarship in a variety of fields. One of the most influential examples is *Archive Fever* (1996), in which Jacques Derrida challenges the concept of an archive as a definable entity with estimable value and an uncomplicated relationship to history and memory. In the seminal essay collection *Refiguring the Archive* (Hamilton et al., 2002), contributors ranging from Derrida to Verne Harris (Nelson Mandela's archivist) to Achille Mbembe (historian and postcolonial theorist) debate the relationship of archives to memory in the context of South Africa's social, cultural, and political history. Archivists such as Michelle Light and Tom Hyry have argued for greater transparency on the part of archivists and an acknowledgement of the subjectivity inherent in organizational and descriptive practices (2002). And scholars like Ann Cvetkovich have articulated a broader understanding of the concept of an "archive," beyond the types of records and other materials found in conventional archives, to include non-traditional, and often more ephemeral, representations of things like memories, feelings, and lived experiences. *Writing in An Archive of Feelings* (2003) about cultural spaces constructed around sex, feeling, and trauma, Cvetkovich laments that their "lack of a conventional archive so often makes them seem not to exist."

So how might forensic technology and its affordances impact the ways in which creators, archivists, and scholars perceive what information is buried or accorded cultural value and whether and how to describe it? How might computers, as complete material and creative environments, make it possible for an individual (or a group) to generate an archives that preserves the ephemeral, the transformative, the everyday, the personal, the painful, and much more, in a variety of audio, visual, and textual genres? My exploration of these and other questions will incorporate the work of the cultural and literary theorists mentioned above, as well as more traditional archival texts and definitions. This paper will consider the ways in which the computer as an archival object challenges notions about the role of archives, the concept of the "archive," and the work of archivists in global contemporary cultures. I will pay particular attention to the ways in which global disparities in access to technology risk creating a future archive that in many ways resembles the colonial archive of the past.

References

(2005). 'Archives'. *A Glossary of Archival and Records Terminology*. Society of American Archivists. `http://www.archivists.org/glossary/` (accessed 9 November 2009).

Cvetkovich, A. (2003). *An Archive of Feelings: Trauma, Sexuality, and Lesbian Public Cultures*. Durham: Duke University Press.

Derrida, J. (1996). *Archive Fever: A Freudian Impression*. Chicago: University of Chicago Press.

Hamilton, C., Harris, V., Taylor, J., Pickover, M., Reid, G., Saleh, R. (eds.) (2002). *Refiguring the Archive*. Cape Town: David Philip Publishers.

Kirschenbaum, M., et al. (2009a). 'Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use'. NEH Office of Digital Humanities. `http://www.neh.gov/ODH/Default.aspx?tabid=111&id=37` (accessed 4 March 2010).

Kirschenbaum, M., et al. (2009b). *Digital Materiality: Preserving Access to Computers as Complete Environments*. San Francisco, CA, 5-6 October 2009.

Light, M., Hyry, T. (2002). 'Colophons and annotations: New directions for the finding aid'. *American Archivist*. **65(2)**: 216-230.