

White Paper Report

Report ID: 98537

Application Number: HD5034608

Project Director: Matthew Kirschenbaum (mgk@umd.edu)

Institution: University of Maryland, College Park

Reporting Period: 9/1/2008-3/31/2009

Report Due: 6/30/2009

Date Submitted: 9/15/2010



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

**White Paper to the NEH Office of Digital Humanities
Level 1 Digital Humanities Start-Up Grant**

May 2009

**Authors: Matthew G. Kirschenbaum (Project Director), Erika Farr, Kari M. Kraus,
Naomi L. Nelson, Catherine Stollar Peters, Gabriela Redwine, Doug Reside**

Contact: mgk@umd.edu

Digital Humanities Level 1 Start-Up funding (\$11,708) was received in support of a series of site visits and planning meetings for personnel working with the born-digital components of three significant collections of literary material: the Salman Rushdie papers at Emory University's Manuscripts, Archives, and Rare Books Library (MARBL), the Michael Joyce Papers (and other collections) at the Harry Ransom Humanities Research Center at The University of Texas at Austin, and the Deena Larsen Collection at the Maryland Institute for Technology in the Humanities (MITH) at the University of Maryland. The meetings and site visits were undertaken with the two-fold objective of exchanging knowledge amongst the still relatively small community of practitioners engaged in such efforts, and facilitating the preparation of a larger collaborative project proposal aimed at preserving and accessing the born-digital documents and records of contemporary authorship.

The grant period was September 2008-March 2009. The only specified deliverable was this white paper; however, as the Outcomes and Next Steps sections (below) suggest, a small initial investment by NEH has yielded significant benefit in the form of infrastructure, knowledge sharing, and future collaboration.



Table of Contents

I. Background	3
II. Institutional and Individual Participants	4
III. Maryland Institute for Technology in the Humanities (MITH), University of Maryland (September 26, 2008)	5
IV. Harry Ransom Center, University of Texas at Austin (November 12-15, 2008)	10
IVa. Meeting between NEH group and representatives from the Bodleian (November 13, 2008)	14
IVb. Flair Symposium: “Creating a Usable Past” (November 13-15, 2008)	15
V. Emory University Libraries, Emory University (December 5-6, 2008)	16
VI. Outcomes	20
VII. Next Steps	25
Appendices	
Appendix 1. Meeting Agendas	27
Appendix 2. Wordle Made From Notes Generated at the Project Meetings	31
Appendix 3. Participant Biographies	32



I. Background

Today nearly all published poetry, fiction, and drama is *born-digital* in the sense that the text is composed with a word processor, saved on a hard drive (or other computer storage media), and accessed as part of a computer operating system. True, some writers will still employ longhand or even mechanical typewriters as a step in their composition process, but sooner or later the text will be keyed into a computer, almost always to be further revised. Often the text is emailed to an editor, along with other correspondence. Editors edit electronically, inserting suggestions and revisions and emailing the file back to the author for approval. Publishers use electronic typesetting and layout tools, and only at the very end of this process is the electronic text of the manuscript (by now the object of countless transmissions and transformations) produced as the static material artifact that is a printed book. This new technological fact about writing is already having an impact, from office work to government and the academy to literature and the creative arts. President Obama's use of a Blackberry and the implications for the Presidential Records Act is a high-profile example of how the public is coming to terms with the consequences of born-digital authorship. In the particular realm of literature and literary scholarship, this means that writers working today will not and *cannot* be studied in the future in the same way as writers of the past, since the basic material evidence of their authorial activity—manuscripts and drafts, working notes, correspondence, journals—is, like all textual production, increasingly migrating to the electronic realm.

Meanwhile, librarians and archivists have not failed to take note as computer storage media, as well as entire computers, have begun arriving on their doorstep as an increasingly routine part of the acquisition of an author's "papers." Notable authors represented with at least some born-digital material in the collections at either Emory or the Ransom Center (the two major institutional repositories in the current study) include Russell Banks, Samuel Beckett, Lee Blessing, John Crowley, Robert De Niro, Michael Joyce, Thomas Kinsella, Bernard Kops, Norman Mailer, Terrence McNally, Tim O'Brien, Salman Rushdie, Ronald Sukenick, Leon Uris, Alice Walker, and Arnold Wesker. Anecdotal evidence from our conversations with archivists elsewhere suggests other institutions have likewise begun accumulating born-digital fonds from some very significant figures. And certainly this phenomenon is only going to grow more prevalent (John Updike, for example, who died during the writing of this report, is known to have been using personal computers since the mid-1980s; likewise David Foster Wallace routinely worked on a computer).

To date, however, the activity associated with processing such born-digital material has been localized and idiosyncratic, and, at least in the US, without much cross-communication among the different archives and repositories involved; moreover, the archives and repositories, for their part, have not yet addressed these challenges with the scholars who will seek to access born-digital literary material in the years to come. Literary scholars are going to need to play a role in decisions about what kind of data survives and in what form, much as bibliographers and editors have long been advocates in traditional libraries settings, where they have opposed policies that



tamper with bindings, dust jackets, and other important kinds of material evidence.¹ This start-up grant therefore brought together scholars, archivists, digital curators, and technical personnel associated with three significant born-digital collections for a series of targeted site visits and planning meetings at each of their respective institutions, with the goal of working towards a larger project proposal designed to address the needs of both archivists and scholars in this new milieu.

II. Institutional and Individual Participants

The Maryland Institute for Technology in the Humanities (MITH) at the University of Maryland (the lead institution) is an established digital humanities center and home to the Deena Larsen Collection, as well as other significant research activity in born-digital preservation and archives; the Harry Ransom Center at The University of Texas at Austin is home to a number of collections with significant born-digital components, including the Michael Joyce Papers (Joyce is generally considered the author of the first hypertext novel, *afternoon* [1987]); finally, Emory University's Manuscript, Archives, and Rare Books Library is home to the Salman Rushdie Papers, which include several laptops and other born-digital media. The nature of each institution's collections and associated research activity is further described in the "Local Contexts" portion of each meeting report, below. We may note at the outset, however, that the digital papers surveyed in this document contain both literary works intended to be experienced on the screen, such as Deena Larsen's *Marble Springs*, as well as those meant to be encountered on the page, such as Salman Rushdie's *Midnight's Children*. For published works, the original form of output, either print or electronic, has potential implications for the kind of research services and tools that the institutions might provide for their users.

In addition to the three primary institutions, during the course of the project participants had significant contact with digital archivists or other specialists at the Library of Congress, Stanford University, the University of Maine, Yale University (Beinecke), the New York Public Library, the British Library, and the University of Oxford (Bodleian). Several of these contacts and opportunities are described in the narrative that follows.

In terms of personnel, we created a very strong interdisciplinary team uniquely situated and qualified to address the challenges presented by this material. Matthew Kirschenbaum, the Project Director (Maryland), has several highly visible publications related to the central concerns of the project, including a cover story entitled "Hamlet.doc: Literature in a Digital Age" in the August 17, 2007 issue of the *Chronicle of Higher Education*.² As an Associate Professor

¹ See G. Thomas Tanselle, "Bibliographers and the Library," *Library Trends* 25 (4) Spring 1977: Trends in the Scholarly Use of Library Resources: 745-61.

² Matthew Kirschenbaum, "Hamlet.doc: Literature in a Digital Age," *Chronicle of Higher Education*, August 17, 2007: B8-9. Also available online: <http://chronicle.com/free/v53/i50/50b00801.htm>.



in the English department, he brings his expertise in textual scholarship to bear on the objects and artifacts of our new digital culture. He was the first academic researcher to access the Michael Joyce Papers at the Ransom Center, and the results of his research are documented in his book, *Mechanisms: New Media and the Forensic Imagination* (MIT Press, 2008). His interests are well-complemented by those of the other project members, who collectively bring credentials in literary studies, information science, copyright and intellectual property, digital preservation, archival science, and textual editing. Erika Farr (Emory), Naomi Nelson (Emory), Catherine Stollar Peters (University at Albany, formerly of the Harry Ransom Center), and Gabriela Redwine (Harry Ransom Center) are all active information professionals presenting papers and running workshops at major conferences on topics directly relevant to the work discussed here. Doug Reside (Maryland) brings advanced technical knowledge to the project as well as a keen interest in textual studies. Kari Kraus (Maryland) combines expertise in textual scholarship and information science with an understanding of intellectual property.

III. Maryland Institute for Technology in the Humanities (MITH), University of Maryland (September 26, 2008)

Local Contexts

In May of 2007, MITH acquired a substantial collection of vintage hardware, software, and other collectible material from the author and critic Deena Larsen. Unlike the Harry Ransom Center or Emory's MARBL, the Maryland Institute for Technology in the Humanities (MITH) is neither a library nor an archive: it is a working digital humanities center. This brings with it certain obvious limitations, but also unique advantages. Founded in 1999 with the aid of an NEH Challenge grant, MITH is the University of Maryland's hub for the theory and practice of digital humanities, cyberculture, and new media, as well as the institutional home of the international Electronic Literature Organization. MITH is thus conceived precisely as an interface between the scholarly and technical communities, a perspective that we think is essential to the current project. At the same time, MITH's institutional situation, encompassing everything from location and physical security to sustainable integration with library resources, creates challenges for ensuring the safety and longevity of an in-house archive. At present, the physical components of the Larsen collection are housed in dedicated (and locked) display cases in MITH's public conference room. Much of the data has been imaged (copied) from the original media, and is stored on a protected server (a so-called "dark archive"). But a number of critical tasks remain, chiefly in the realm of the creation of metadata and finding aids to turn the archive into a viable scholarly collection.

While not a household name in wider literary circles, Larsen has been an active member of the creative electronic writing community since its inception in the mid-1980s. She is an avid collector and amateur archivist (or hoarder) who was happy to find a home for the dozen or so vintage Mac Classics, roughly 1000 diskettes, and boxes of journals, papers, correspondence, newspaper clippings, memorabilia, and ephemera previously stored in her apartment. In addition



to her own writing and creative work, Larsen also possesses a broad array of material by other electronic literature authors, some of it unpublished, unavailable, or believed otherwise lost. MITH, for its part, looks upon its acquisition of the Larsen collection as both an important service to electronic literature (by safeguarding what Larsen herself has described as that community’s “great library of Alexandria”) as well as an invaluable research opportunity, given the potential of this material to function as a testbed.



Figure 1. The Larsen Collection at MITH

The Larsen collection is part of a broader spectrum of digital preservation activity at MITH. Assistant Director Doug Reside is playing a key role in evaluating the Jonathan Larson digital papers for the Library of Congress. (Jonathan Larson—no relation to Deena—is best known as the composer of the musical *RENT*.) In December 2008, Kirschenbaum and Reside collaborated with the university’s Digital Forensics Lab on the recovery and emulation of William Gibson’s famous electronic poem *AGRIPPA*, an achievement covered on Slashdot, Boing Boing, Wired, and numerous other venues around the Web.³ Finally, Kirschenbaum, Reside, and MITH Director Neil Fraistat are all participants in a Library of Congress NDIIPP-funded project on Preserving Virtual Worlds.⁴ The acquisition of the Larsen collection therefore fits into a strong emerging research profile in digital preservation at MITH.

³ See <http://agrippa.english.ucsb.edu/kirschenbaum-matthew-g-with-doug-reside-and-alan-liu-no-round-trip-two-new-primary-sources-for-agrippa>.

⁴ <http://pvw.illinois.edu/pvw/>.



Last but not least, MITH enjoys a working relationship with the university's Information School (iSchool), where archival science is an area of recognized strength. (Allen Weinstein, Ninth Archivist of the United States, has just joined the faculty.) Kari Kraus, an assistant professor with a joint appointment in the iSchool and English department with particular expertise in copyright and the intersection of textual scholarship and archival theory, is a member of the current project, as well as the aforementioned Preserving Virtual Worlds, and Rachel Donahue (currently an MLS student, about to begin the doctoral program at Maryland) is a graduate assistant with Preserving Virtual Worlds and also participated in the local meeting for this project. The iSchool therefore provides MITH with access to important additional expertise, also notably in the person of senior archivist Bruce Ambacher on its faculty.

Meeting Summary

Participants: Erika Farr, Matthew Kirschenbaum, Kari Kraus, Naomi Nelson, Gabriela Redwine, Doug Reside, Catherine Stollar Peters. Also present: Rachel Donahue (Maryland, College of Information Studies, doctoral student in archives).

For the first meeting in the series, we had the joint goals of initiating conversation on a range of issues likely to be relevant throughout the project and introducing participants to the specifics of the collection and environment at Maryland. We broke the ice with a casual dinner on the evening of September 25th at the College Park home of principal investigator Matthew Kirschenbaum.

Work began in earnest on September 26th with an all-day meeting at MITH. Following a format that would also be loosely adopted for each of the two subsequent meetings, we devoted the morning to an in-depth consideration of the particular born-digital collection(s) at the host institution, enjoyed a working lunch with one or more special guests, and then spent the afternoon with additional guest presentations and covering related work. The day ended with a brainstorming exercise designed to recap the day's discussion and lay the groundwork for the next meeting in the series. The original agenda is available as Appendix A.

During his overview of the Deena Larsen collection, Kirschenbaum covered much of the background already provided in the previous section on Local Contexts. One point of particular emphasis was the hybrid nature of the Larsen materials. Her most significant work, *Marble Springs* (1993), exists in a number of physical and digital states which exhibit complex relationships and dependencies. A shower curtain, for example, is the support for a dozen laminated screenshots representing different pieces of the work; these are connected by colored yarn mapping their links and relations. An artifact such as this, coupled with hard copy print outs and transcripts, coupled with digital drafts in various formats and versions of the HyperCard software used as the final authoring environment, is emblematic of the kind of challenge archivists in a number of different cultural heritage sectors can expect to face in the future: not just born-digital content, but digital-analog hybrids. Following Kirschenbaum's presentation, we



devoted some hands-on time to allowing participants to inspect the actual content of the collection.

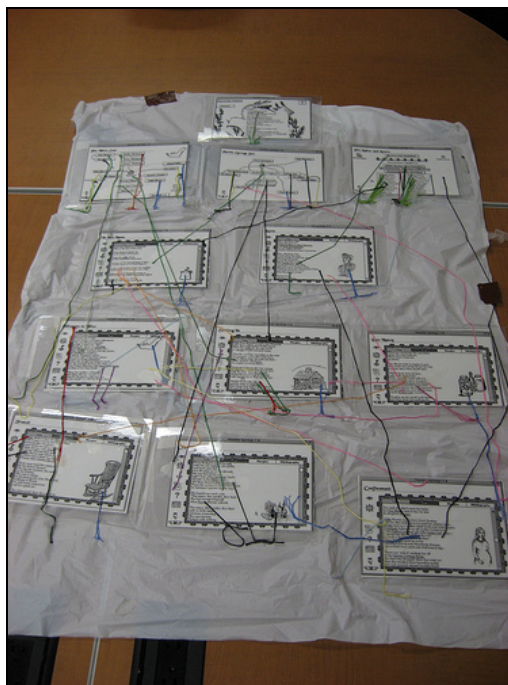


Figure 2. The Larsen Shower Curtain

The highlight of the day was perhaps our conversation with two local writers who joined the group for lunch. Elizabeth Arnold is a poet on the faculty of the University of Maryland, with two published books and a number of prestigious awards and fellowships to her credit. William Loizeaux is a local independent writer who has authored children's books and non-fiction. We asked the writers to join us because there is scant documentation of how word processing and other digital technologies are changing working authors' actual writing habits. We were interested in the most mundane details, such as whether composition begins at the keyboard for these writers or if they work with pen and paper, how often they save revisions and versions, what their backup strategies are, whether they have a Web browser open while they write, how they handle their email, and whether and how they think about the privacy issues that would arise with an archival examination of their computer. We were particularly struck by the question of whether it was ethically appropriate for an archivist to intervene early on in a writer's career—to encourage best practices related to the long-term stability of the born-digital materials—because of the way in which this might impinge on an author's creative process. The conversation with Arnold and Loizeaux clearly established the importance of further communication with working authors, and we spent some time discussing practical strategies for achieving this, such as a future presentation at the Association of Writing Professionals conference.

After lunch with the writers, Kirschenbaum, Kraus, and Reside offered overviews of the related work with which they are involved. Reside discussed his work with the Jonathan Larson materials at the Library of Congress, where his efforts have already saved important cultural records. The Larson materials present challenges related not only to textual analysis but also obsolescent audio formats. Kirschenbaum and Kraus discussed their role in the aforementioned NDIIPP-funded Preserving Virtual Worlds project, whose scope ranges from interactive fiction (so-called “text adventures”) produced in the 1970s to the vast and persistent multi-user world of Second Life. Kraus devoted particular attention to intellectual property issues, such as the slippage that occurs between copyright and contract law when one is dealing with software licensing; also the role of the amateur archivists (the hobbyist or enthusiast) who work to preserve content such as video games. Finally, Kirschenbaum discussed the Variable Media Network, whose research has been funded by the NEH under the titles Archiving the Avant Garde and Forging the Future. (Kirschenbaum acts as a consultant with this group.) Based in the museum rather than the library community, the Variable Media Network has produced an interlinked set of cataloging tools devoted to documenting the behaviors and providing guidelines for the reconstruction of works of digital and performance/installation art. Their approach has resulted in MANS, the Media Art Notation System, an adaptation of MPEG-21 for encoding the behaviors of complex interactive work.⁵

The final activity of the day was a discussion of conclusions and next steps, based on an exercise in which each group member brainstormed ideas and wrote them on Post-it notes which were then grouped into related clusters on a whiteboard. This activity revealed a wide ranging and fertile discussion that opened a number of avenues for future exploration, while also (already) suggesting the need for priorities and focus in any future project the group might undertake. The issues discussed included hardware and physical storage, in particular the relationship between archives and computer museums and the inevitable deterioration of physical media (so-called “bit rot”). Metadata and standards also received attention, including METS, PREMIS, FRBR, and the OAIS Reference Model. We discussed the role of emulation in providing access to a computer as a complete environment, and how this differs from storing files in a repository such as Fedora or DSpace. A large amount of attention focused around issues of privacy and confidentiality: how does one balance reasonable expectations of donor privacy against the need for researcher access? (What if there were multiple users of a particular machine?) Discussion also encompassed intellectual property issues, notably the status of pirated software that finds its way into a repository by means of an acquisition such as the Deena Larsen collection. Cloud computing and its implications for preservation were acknowledged as a major challenge: the task is hard enough when the physical media are there in front of you, but how does the archivist contend with the increasingly ubiquitous ecology of Web 2.0 applications, each of which is governed by a different End User License Agreement? We discussed the role of contingency and

⁵ For MANS, see Richard Rinehart, “The Media Art Notation System: Documenting and Preserving Digital/Media Art,” *Leonardo* 40, No. 2, (2007) pp. 181–187.



circumstance in preservation efforts to date: for example, someone who just happens to discover the right legacy component in a box in a supply closet, without which vital content on obsolete media might have remained inaccessible. Is there a way to build a knowledge base of resources that could be made available across a number of different institutions so that multiple repositories could use this same legacy component to rescue a greater number of files? Finally, the conversation returned to our luncheon with the writers, and how to balance among the needs of author, archivist, and scholar; we discussed at what point to involve the author in the archival process, and likewise, what role scholars should play. This last point was perhaps the most essential insight to emerge from the initial meeting: namely, recognizing the triad of stakeholders who need to play a part in formulating approaches to managing and collecting born-digital literary materials for scholarly use.

At the conclusion of the meeting these Post-its were collected and transcribed for future reference. The group adjourned to dinner, where conversations continued. There was unanimous consensus that the initial meeting had been productive, informative, and intellectually exciting.

IV. Harry Ransom Center, University of Texas at Austin (November 12-13, 2008)

Local Contexts

The Harry Ransom Center is a literary collecting archive whose primary emphasis is the study of the literature and culture of the United States, Great Britain, and France. In addition to its extensive manuscript, book, photograph, art, and film holdings, the Center also houses the computers and disks of authors such as Norman Mailer, Arnold Wesker, Michael Joyce, and Terrence McNally. Although the Ransom Center has been receiving born-digital items as part of paper collections for over a decade, the 2005 Michael Joyce acquisition marked the Center's first deliberate engagement with born-digital literary materials published in electronic format.

Archivists at the Ransom Center, in collaboration with students from the School of Information (iSchool) at the University of Texas at Austin, have been processing the Joyce materials in stages since 2005. The bulk of the digital collection was processed by a group of three students in the iSchool master's program. This initial project marked the beginning of the Ransom Center's collaboration with the iSchool, and in particular with Dr. Patricia Galloway and her graduate students, on digital preservation projects. In the last three years, the Ransom Center has worked with Dr. Galloway and her students to catalog the majority of the Joyce materials, as well as all of the disks in the Leon Uris, John Crowley, and Arnold Wesker holdings, and portions of the digital materials in the Norman Mailer and Terrence McNally papers.

The Ransom Center currently has 35 holdings that contain electronic records. These materials include born-digital items received with manuscript collections, as well as software, some hardware, and manuals. During accessioning, electronic records are transferred to the Electronic Records Collection, and the accessioner notes those transfers both in the Manuscripts Accessions and the Manuscripts Collection databases. The digital archivist records disk-level metadata and



collection information into an inventory of all of the Ransom Center's electronic records. A new database with collection-level records to describe the content, status, and access available for each electronic records collection is now in progress. The Center is also in the process of updating its policies and procedures for born-digital materials and writing a processing manual for staff and students to use as they catalog these materials.

The Ransom Center currently stores its digital materials in a DSpace repository hosted by the iSchool. Access to the files in DSpace is restricted to students working on projects, Center staff, and iSchool DSpace administrators. The Ransom Center offers access to its digital materials on a case-by-case basis. In the last two years, three patrons have requested access and the Center has been able to accommodate all of them. Two of the patrons used the Arnold Wesker and Michael Joyce materials via DSpace in the reading room; the third patron accessed copies of files from Terrence McNally's disks directly from a laptop in the reading room. In addition, the Center's electronic records collection has been represented in two in-house exhibitions: *Technologies of Writing* (2006) and *The Mystique of the Archive* (2008).



Figure 3. Items from the Michael Joyce Papers on display at the Ransom Center as part of the *Mystique of the Archive* exhibition

The Ransom Center's work with digital materials is innovative in a few key ways. First, the Center's approach has been holistic: the digital archivist not only preserves the computer files but also organizes and describes the entirety of a collection by integrating paper and digital materials, makes those files available to patrons, and incorporates them into exhibits. Of the repositories participating in the NEH grant, the Ransom Center has the most experience working



with digital collections: it holds the largest number of collections with electronic records, has created procedures for arranging, describing and using electronic records, and has an established OAIS-based repository for preservation and use. However, the Ransom Center has focused preservation efforts purely at the file level and series level, and has undertaken little research on preserving disk images—something the digital archivist hopes to correct in the near future.

Second, the Center's productive collaborative relationship with iSchool students and faculty is unique. That relationship has enabled the Center to rely on the assistance of iSchool students to process the digital materials in several collections and in many ways serves as a model for other institutions. Working with students also helps Ransom Center archivists keep abreast of recent developments, scholarship, and technological advances relevant to the field of digital preservation and, in turn, gives students real-world experience to bring to their future employers.

Third, the Ransom Center has begun to solicit feedback from scholars. The digital archivist designed a questionnaire asking patrons to evaluate their experience using digital collection material in the reading room, and has administered it to two visiting Fellows. This feedback will help the Center: (1) improve the level of access to digital items and (2) ensure that the reference services and technology offered to scholars who come to use authors' computer files matches the high standards of the service provided for traditional archival material. A related outcome, which was a topic of conversation at the third site visit, at Emory, is a growing awareness of the difference between the scholarly and archival perspectives when it comes to thinking about how best to manage digital collection material. This type of feedback from scholars, combined with input from archivists, will be immensely helpful in shaping future research into what types of access tools manuscript repositories might want to provide.

Although there is a variety of digital work being done at the University of Texas, the archivists, librarians, technologists, and scholars working with digital materials there do not yet form a cohesive community. In addition to the Ransom Center, one other archival repository on campus—the Center for American History—employs a digital archivist. A few of the other repositories on campus have born-digital collection material but do not currently have the personnel or funding resources to support digital preservation work on those items.

Meeting Summary

Participants: Erika Farr, Matthew Kirschenbaum, Kari Kraus, Naomi Nelson, Gabriela Redwine, Doug Reside, Catherine Stollar Peters. Also present: Michael Forstrom (Beinecke), Neil Fraistat (MITH), Tom Hyry (Beinecke).

The purpose of the Austin site visit was to introduce the Maryland and Emory grant participants to the Ransom Center's digital collection material, use the Michael Joyce and Arnold Wesker projects to illustrate the Center's basic strategy for processing computer files discovered in otherwise paper collections, give meeting attendees a chance to experience hands-on what type of access to digital materials the Ransom Center is able to provide to reading room patrons, and



demonstrate some of the ways in which the Center has incorporated digital manuscripts and related artifacts into exhibits. The group scheduled this Nov. 12th meeting to coincide with the Flair Symposium being held at the Ransom Center from Nov. 13-15, in order to take advantage of the convergence of digital archivists, scholars, librarians, and others and continue our conversations over an entire week. This timing also enabled us to include two additional participants—Michael Forstrom (Archivist) and Tom Hyry (Head of the Manuscript Unit) from the Beinecke—both of whom were in town for Flair. Additional benefits derived from the Flair Symposium will be discussed in more detail below.

To give participants a sense of the promise and challenge of our digital collection materials, organizers Redwine and Peters created two different exhibitions that meeting attendees were able to access throughout the day. The first consisted of a set of electronic collection materials installed on three networked computer workstations around the room. The point of these workstations was to give people an idea of what a patron would experience upon visiting the Ransom Center's reading room to look at digital manuscripts and correspondence. Attendees looked at files from four different collections and accessed them both from the desktop and through DSpace, the Ransom Center's online repository. These files included different versions of some of Michael Joyce's manuscripts, as well as digital materials from the Terrence McNally, Arnold Wesker, and Tom Zigal collections. One of these items, which highlights the intersection of creation and technology and the effect of technology on an author's work, is a stream-of-consciousness document McNally typed on 10 June 1988 as he experimented with WordPerfect for the first time. Redwine and Peters also included a set of proofs (created in Word) that Tom Zigal exchanged with his editor at The Toby Press. Their tracked changes and comments provide valuable insight into the creative process. Both sets of materials offer a precise illustration of the complex motivations for this grant: to understand and preserve authors' works and the environment in which they are created.

The second was a small exhibition of disks and computers from the Center's collection. Peters and Redwine incorporated these items and their respective histories into an introductory overview to give attendees a sense of some of the particular digital preservation challenges presented by the media formats in the Ransom Center's collection. In addition, items like a letter from author Bernard Kops in response to Redwine's query about his computer usage illustrated the importance and utility of beginning conversations with living authors about their technology habits.

After preliminaries, Redwine showed the group the bank of "legacy computers"—a series of older Mac and PC desktops with different types of software and disk drives—that archivists and students use to process digital materials at the Ransom Center. The questions that arose about these computers and the policies and procedures regarding digital materials led to a general discussion about the Center's collections and digital processing practices. Afterward, Peters talked about her experiences with processing Michael Joyce's laptop and disks; attendees were able to use the computer workstations to access some of Michael Joyce's files during and after



the presentation. Next, Dr. Patricia Galloway presented on her current research projects and the history of the digital preservation program at the School of Information. Both Redwine and Peters are former students of Dr. Galloway, and her presentation shed light on the development of the Ransom Center's approach to preserving born-digital materials. The presentation also highlighted future research projects in digital preservation that would rely on techniques from the fields of digital forensics and data processing.

Over lunch, Zach Vowell from the Center for American History talked about their Videogame Archive, the types of materials he has come across, and the work he has done on it so far. Vowell is the only other digital archivist on the University of Texas at Austin campus, and his repository also relies on iSchool students to help with digital processing projects. Peters opened the afternoon session with a presentation on the scope and challenges of the computer files in the Arnold Wesker collection, and Redwine discussed the lessons she had learned from providing a Ransom Center Fellow with access to Wesker's digital materials in the reading room this past summer. This session was instrumental in defining some of the potential areas for further research involving user access to born-digital materials. Carlos Ovalle, a copyright and intellectual property specialist from the School of Information, rounded out the day with a brief presentation about issues specific to digital manuscript collections, which led to a group question and discussion period. Carlos's expertise in copyright answered a number of questions about providing access to copyrighted materials within and external to the "four walls" of an institution.

The last hour or so of the meeting was devoted to participating in the same Post-it brainstorming exercise, discussing what the group had learned at the Ransom Center, and talking further about some of the issues and questions common to the Austin and Maryland site visits. We also laid general plans for our next NEH grant meeting, which took place on December 5th at Emory. And, finally, we discussed the next day's meeting with Richard Ovenden and his colleagues from the Bodleian.

IVa. Meeting between NEH group and representatives from the Bodleian (November 13, 2008)

Participants

Bodleian Library: Chris Fletcher (Head of Western Manuscripts), Renhart Gittens (Software Engineer, futureArch project), Richard Ovenden (Head of Special Collections and Associate Director), Susan Thomas (Digital Archivist and futureArch project manager).

Harry Ransom Center: Gabriela Redwine (Archivist and Electronic Records Specialist), Catherine Stollar Peters (former Archivist and Electronic Records Specialist [now at New York State Archives]).

Emory University: Erika Farr (Director of Born-digital Initiatives, Robert W. Woodruff Library),



Naomi Nelson (Interim Director, Manuscript, Archives and Rare Book Library).

University of Maryland: Neil Fraistat (Professor of English and Director of MITH), Matthew Kirschenbaum (Associate Professor of English and Associate Director of MITH), Kari Kraus (Assistant Professor, iSchool, College of Information Studies), Doug Reside (Assistant Director of MITH and a Visiting Assistant Professor of Theater).

The purpose of this meeting was to learn more about the digital preservation work happening at each institution and discuss possibilities for future collaboration. A representative from each repository presented on a project or a digital collection, and then we talked more broadly about what each of us has done with our digital materials thus far. Matthew Kirschenbaum summarized the NEH project work to date, including the previous day's discussions. Gabriela Redwine from the Ransom Center spoke about the Michael Joyce Papers, and then gave a quick overview of the extent of the Center's digital holdings and their processing status. Susan Thomas from the Bodleian talked about the futureArch project, which is geared toward implementing digital preservation infrastructure and developing researcher interfaces for hybrid archives. Erika Farr from Emory talked about the Salman Rushdie computers, and Matthew Kirschenbaum and Doug Reside from MITH discussed the papers of Deena Larsen and Jonathan Larson (*RENT*).

Since this meeting was exploratory, our outcomes reflect a commitment to sharing information and resources. We agreed to the following two “action items”:

1. Create a restricted online space we can use to notify each other about resources, share information about some of the hardware and software in our collections, and discuss the challenges we encounter during processing. Redwine created this space (a Google Group), and we have already begun using it to share information.
2. Consider applying for a grant to fund a two-part symposium at one repository in the U.S. and another in England. Each part of the proposed symposium would span two days. Day one would be open to anyone interested in the topic of digital preservation, and the goal of that part of the symposium would be education and outreach among archivists, iSchool faculty and students, patrons, digital curators, and other interested parties. Day two would be a series of focused presentations and discussions limited to digital archivists and scholars from a select group of institutions, and would also include a site visit component. (A modified version of this is currently pending as a proposal at a major foundation.)

IVb. Flair Symposium: “Creating a Usable Past” (November 13-15, 2008)

We decided to hold the NEH site visit at the Ransom Center in conjunction with the Center's biennial Fleur Cowles Flair Symposium (Nov. 13-15), in part because this year's Flair theme, “Creating a Usable Past: Writers, Archives, and Institutions,” spoke directly to the questions underlying our grant. The symposium not only drew an impressive and international array of archivists, technologists, digital archivists, digital curators, dealers, agents, and scholars, but



also, importantly, included writers. During the symposium, two separate panels composed of Lee Blessing, Denis Johnson, Tim O'Brien, and Amy Tan addressed the relationship between archives and author, offering a rare opportunity for authors to speak to one of the recurring themes in the NEH and Bodleian Library meetings: How do authors manage their electronic and paper materials before archiving them at an institution? NEH grant participants were able to hear first-hand how authors view their materials, laying a foundation for future approaches to preservation and highlighting the need for more interaction and understanding of the creative practices relating to electronic records. On the morning of the 15th, Kirschenbaum and Nelson participated on a panel devoted to the (digital) future of archives with Susan Thomas from the Bodleian, as well as Peter Hirtle (Cornell) and Jeremy John (British Library), moderated by Kris Kiesling (Minnesota).

In many ways, the Flair Symposium was a continuation of the conversation begun at the Ransom Center site visit on the 12th and the meeting with the Bodleian representatives on the 13th. We used the additional days to strengthen contacts with other archivists and scholars, discuss our work on born-digital literary papers, and learn about some of the digital preservation and curation projects happening at repositories in the UK. As the result of those conversations, several of our group members travelled to London in February for the Digital Lives Conference at the British Library and to Chapel Hill, NC, at the end of March for the “Stewardship of E-Manuscripts: Advancing a Shared Agenda” meeting organized by Tom Hyry (Beinecke) and Cal Lee (University of North Carolina).

V. Emory University Libraries, Emory University (December 5-6, 2008)

Local Contexts

The Emory University Libraries (EUL) include numerous branches and divisions, in particular the Manuscript, Archives, and Rare Book Library (MARBL) and Digital Systems. Representatives from these two components of the EUL participated in this NEH planning grant. The Emory University Libraries are engaged in digital libraries, digital humanities, and scholarly communication and have pursued innovations in these areas for over a decade. MARBL's acquisition in 2006 of Salman Rushdie's personal papers provided it with a rich personal archive of historical and literary significance that includes analog and digital artifacts. This acquisition is not only important to Emory's growing collection of twentieth century literary papers but also to the emerging strategic initiatives around born-digital content and archives.





Figure 4. The Rushdie computers at Emory

The Rushdie archive marks MARBL's first acquisition of a significant amount of born-digital material and includes four personal computers and one external hard drive. With the arrival of this hybrid archive, MARBL and Digital Systems formed a working group consisting of archivists, preservation experts, hardware specialists, and digital librarians. The Born-Digital Archives Working Group (BoDA) developed a preliminary project plan for undertaking the acquisition, processing, and preservation of the digital material. In addition, BoDA explores optimal approaches to organizing and presenting these materials and their analog counterparts as a seamless hybrid archive for research purposes. Archivists completed the arrangement and description of Rushdie's analog records in February 2009. As for the born-digital records, they have created disk images for all computers and drives acquired as part of this archive and indexed these drives. Staff is also currently identifying duplicate files across the archive and assessing the breadth and depth of born-digital material included. Over the next two years, MARBL, with the support of BoDA, intends to provide research access to this important collection of literary papers in stages, with limited access to manuscript and manuscript-related digital files as the first stage, and a later final stage of full access to the paper and digital records through a range of tools including fully searchable text databases as well as system and application emulation packages.

Meeting Summary

Participants: Erika Farr, Matthew Kirschenbaum, Kari Kraus, Naomi Nelson, Gabriela Redwine, Doug Reside, Catherine Stollar Peters. Also present: Laura Carroll, Susan McDonald, Lars Meyer, Peter Hornsby (all Emory University Libraries).



As the last event in this series of planning meetings, the gathering in Atlanta at Emory University was intended to extend the conversations begun at the Maryland and Austin meetings, synthesize the group's topics of interests, and identify the next steps for the project members.

The meeting began with a kick-off dinner on December 4. This gathering gave the group an opportunity to share information informally and provide updates on various activities at home institutions and in the larger field. Friday, December 5, consisted of a day-long meeting at Emory, with project members and various invited guests in attendance. The first half of the day was largely devoted to a discussion of born-digital initiatives at Emory University, with a significant portion of the morning dedicated to the processing of and prospects for the Rushdie papers. Several members of the Emory University Libraries staff joined the project team for these discussions, including archivists Laura Carroll and Susan McDonald, who are processing the paper records of the Rushdie archive; and Pete Hornsby, the lead technologist working on Rushdie's computers, who presented on Emory's approaches to assessing, stabilizing, and processing the born-digital content. The Emory team discussed the ways in which Rushdie's paper and digital records were similar and different—including the way he organized records and the kinds of records he kept in each format. They also touched on the growing diversity of file formats represented on the machines over time. The electronic files represent Rushdie's entire digital life from 1992-2006. Lars Meyer presented on digital curation and Emory's nascent Digital Curation Center. The project team was particularly intrigued by his description of digital curation as "managing polarities," which he illustrated with the following examples:

- Preserve object and preserve environment
- Preserve functionality and preserve fidelity
- Trust extant metadata and enhance metadata
- Allow use and respect confidentiality
- Allow use and ensure rights compliance
- Appraise and select, and be complete

For lunch, the project team members met with Natasha Trethewey, Pulitzer Prize-winning poet and Emory faculty, to discuss her archiving practices and how she, as both poet and scholar, understands the role of digital media in the formation of literary papers. This conversation was candid, casual, and deeply informative. Topics ranged from her conscious removal of content and artifacts from her personal papers to the intellectual property rights implications of composing poetry on University laptops. In response to questions about how her use of technology may have influenced her creative process, she described how the use of a word processor allowed her to experiment more with the arrangement of her words on a page. Trethewey also discussed her recent purchase of an iPhone and her use of technology in general.



In the analog world, she did not experiment as widely because of the time it took to retype a poem. In many ways, this conversation was an extension of one started in Maryland with the poet Elizabeth Arnold and writer William Loizeaux.

In the second half of the Emory meeting, we invited two guest speakers to join the conversation. Katherine Skinner, an Emory librarian and project manager of the NDIIPP-funded MetaArchive project (<http://metaarchive.org>), shared with the group her findings and experience with distributed digital preservation. Emory was a charter member of this collaborative, community based initiative to build a trusted repository to preserve cultural memory. It is using the open source LOCKSS software to provide distributed digital preservation. Bill Underwood, from the Georgia Tech Research Institute, was the final speaker and provided the group with a wealth of information and innovative ideas on developing applications and tools for processing born-digital records. Underwood is the Principal Investigator for The Presidential Electronic Records Pilot System (PERPOS) project sponsored by the U S National Archives and Records Administration (<http://perpos.gtri.gatech.edu/>). The goals of the project are as follows:

- To analyze and design software tools that support the Accession, Preservation, Arrangement, Review, and Description of electronic records.
- To evaluate experimental and commercial natural language processing (NLP) search and retrieval tools for use in reviewing Freedom of Information Act (FOIA) exceptions, reviews of Privacy Act and Presidential Records Act restrictions, and responding to routine reference and FOIA records requests.
- To refine and implement the data model in an object/relational database for accession, arrangement, description and access.
- To formulate a metadata model and associated XML Schema for storage, preservation, and content access.

The tool and application prototypes Underwood presented sparked lively conversation and highlighted numerous possibilities for future collaboration, particularly in the areas of automated arrangement, description, and redaction.

The final session of the day involved identifying common themes and interests that had emerged during the day and reviewing the agenda for the half-day meeting to follow on Saturday. As had been true in the two preceding meetings, concluding the day with brainstorming and discussion offered an important opportunity to reflect on the day-long conversation and synthesize perspectives and information.

On Saturday, the sixth of December, the project team members from MITH, the Harry Ransom Center, and Emory reconvened to review the progress and outcomes of the three planning meetings, identify desired outcomes for the planning project, and develop a strategy for



accomplishing those outcomes. The Emory hosts had requested that all project partners identify the strengths and opportunities of the three institutions and each individual participant. We began the meeting on Saturday by reviewing these submissions. This exercise provided the group with a broad view of the institutions and individuals involved and offered insights into how and why these groups should collaborate on approaches to collecting and managing born-digital literary papers. The activity also allowed us to discuss candidly the differences in perspectives, training, and skill sets among this diverse group of archivists, technologists, scholars, and librarians. Clear understanding of the priorities and values of partnering institutions and mutual respect among partners are vital to successful collaboration, so this exercise, while time-consuming, was worthwhile.

Next, the group discussed the two desired outcomes of the planning grant: this white paper summarizing the work and outcomes of the grant and a larger grant application. The group quickly agreed upon the format and objectives of the white paper and developed a strategy for accomplishing this work. The larger question of developing a follow-on grant application that could fund collaborative efforts at processing and providing access to born-digital archives led to extensive discussion of what kind of work we would like to pursue with additional funding and how we might structure such a collaborative project. Some ideas that emerged for possible grant work include developing tools for managing born-digital content, creating emulation prototypes, and drafting a workbook or set of guidelines. More immediately, the group agreed to focus on producing presentations and articles from our project findings, as well as continuing to develop a collaborative network of professionals, scholars, archivists, and technologists who are thinking about and working on these topics.

VI. Outcomes

Born-digital preservation and records management are still very young specializations. While some impressive guides to best practice already exist (notably the Paradigm Workbook on Digital Private Papers prepared by staff members at the Bodleian and Rylands [Manchester] libraries),⁶ and while research is under way in certain quarters,⁷ it is clear that the field will remain in a state of flux for the foreseeable future. Many challenges exist for which there is simply not enough accumulated wisdom and experience to formulate best practices. In particular, the question of scholarly access to born-digital collections remains an uncharted wilderness. Of the three institutions involved in the current project, only one (the Ransom Center) has actually implemented an access strategy, and it allows for only item-level access via a DSpace repository

⁶ <http://www.paradigm.ac.uk/workbook/index.html>.

⁷ For example, Jeremy Leighton John, "Adapting Existing Technologies for Digitally Archiving Personal Lives: Digital Forensics, Ancestral Computing, and Evolutionary Perspectives and Tools," 29-30 September 2008 iPRES 2008 Conference, The British Library. Adrian Cunningham's work in this area is especially significant: the earliest paper of which we are aware is Cunningham, "The Archival Management of Personal Records in Electronic Form: Some Suggestions," *Archives and Manuscripts* 22.1 (1994): 94-105.



or access to files copied to a laptop housed in the onsite reading room. Access strategies are particularly vexed in the realm of literary (and other creative) born-digital content because, given the nature of the creative process and the kinds of questions scholars are likely to want to pursue, it is difficult even to achieve consensus on the proper object of preservation. Is it the physical hardware? A mirror of the original operating system, including all systems files, all applications, and the desktop? Bit-level disk images allowing for forensic analysis of overwritten data? And so forth. Indeed there is a very strong case to be made for treating all born-digital literary and creative collections as *special collections*, where multiple facets of the original artifact are likely to have value, and for multiple constituencies. Returning to Tanselle's cautions (see Background, above), every effort must therefore be made to ensure we are not inadvertently discarding the dust jackets of the digital age.

To that end, everyone in the project group shared a keen appreciation for what Richard Ovenden has helpfully called (in conversation) “the digital materiality of digital culture.” We would gloss this as a curatorial sensitivity toward the uniqueness of individual instances of both hardware and data objects, coupled with an awareness of how the affordances of particular systems, environments, and technologies can all impact the creative process. For example, knowing how much of a document would be visible on a screen at one time—knowledge that depends on the physical size of the display hardware, its screen resolution, and preferences as defined within particular application software—can be critical to understanding aspects of an author's composition process. Terrence McNally comments on precisely this phenomenon in the stream of consciousness WordPerfect document mentioned in part IV above. “This is the 22nd line,” he writes. “After I finish it and two more, the screen should begin to move upwards and I will only be seeing the last 25 lines. It is not possible to see an entire document when you work with a computer.”⁸ Umberto Eco had Belbo similarly experiment with his new computer in the novel *Foucault's Pendulum* (see the beginning of chapter 3), and Salman Rushdie has equivalent files on his Macintosh laptops, showing that he, too, took time to explore the environment of his new computer. The experience of an author composing on a Mac Classic from 1985 will be different from the experience of an author working on a contemporary wide-screen LCD display, or perhaps several such displays configured in tandem. Textual scholars have been attentive to the “materiality” of books and manuscripts for decades, especially following the intervention of scholars such as D. F. McKenzie and Jerome McGann who formulated influential approaches to the theory of scholarly editing in the 1980s. As we move forward into the digital era, we would thus do well to monitor the extent to which archival practice is or is not in step with the intellectual convictions of neighboring scholarly communities.

Closely related to these questions of materiality is the hybrid status of nearly all born-digital collections of which we are aware, in which electronic objects coexist with more traditional forms of archival content. Material from collections at all three of our institutions exemplifies

⁸ Terrence McNally Papers, Harry Ransom Humanities Research Center, file name: NEWLIGHT.



this phenomenon, with the textual horizons of a particular work often spanning multiple media and formats, from holograph manuscript to hard copy print-out of a born-digital text, to actual digital files. (The Larsen shower curtain is perhaps the limit case.) Scholars will want and need to track the evolution of a work without regard for the gaps and incompatibilities introduced by competing or obsolescent data formats and operating systems, let alone the analog/digital divide. Yet there are no tools to facilitate this kind of activity, and there are unlikely to be for the foreseeable future. Compounding the problem is the reality that working authors often gravitate toward proprietary software, such as the word processor that came installed with their system as a default. While the digital humanities community has had reasonable success to date in developing text analysis, text mining, and visualization tools for large electronic corpora, these tools often assume ideal circumstances and a homogeneous data set, not the messy world of proprietary and mutually incompatible formats one gets from an individual user's hard drive. In many cases the only exploratory and analytical tools likely to be of use will be the Finder applications built into current operating systems, and there is no reason to assume their availability as part of an access environment. At one end of the spectrum we can anticipate expanding metadata for finding aids from collection level to item level information to more robustly track the migration of a work across multiple media and formats. At the other, more exotic, end of the spectrum it is perhaps possible to imagine grafting RFID tags to physical archival objects in order to convert them to what Bruce Sterling has called "spimes," that is physical objects digitally locatable in space and time, thereby making linkages to associated data explicit.⁹ Regardless of what solutions are actually deployed, it is clear that both archivists and scholars will need to contend with increasingly complicated ecologies of primary source documents spanning heterogeneous digital and analog states.

Throughout the grant period, we encountered a range of issues and challenges which we came to identify as "persistent but non-primary," meaning that while pervasive in our discussions, they themselves alone ought not to constitute the nucleus of our future effort since they do not leverage the unique strengths of our group. These included: the physical maintenance of hardware and storage media; copyright and intellectual property; concerns over privacy, confidentiality, and security; questions about the valuation and appraisal of born-digital objects; and the new and formidable challenges presented by cloud computing, that is the increasing reliance of network-centric services for email, blogging, photo-sharing, and social networking. The growing user base for third-party back-up services like iDisk and Carbonite, and the imminent arrival of Google's remote storage capacity only further exacerbate the extent to which archivists will have to cope with a bewildering array of terms of service. The novelist Zadie Smith, for example, has expressed uncertainty over who actually owns the email she sends with a Yahoo account.¹⁰ (And during the writing of this report, Facebook announced and then abruptly backed away from revised terms of service that sought to claim rights over any content ever

⁹ See Sterling, *Shaping Things* (Cambridge: MIT Press, 2005).

¹⁰ Rachel Donadio, "Literary Letters, Lost in Cyberspace," *New York Times*, 4 September 2005.



uploaded to one of its accounts, even if the user subsequently terminates the account and leaves Facebook.) Even if the ownership of digital content is clear, a corporation's control of access to the content can render ownership meaningless. In 2004, Yahoo cited its privacy policy when it denied the family of a soldier killed in Iraq access to his Yahoo email. The family received access only after obtaining a court order.¹¹ In 2007, LiveJournal (a blog site popular with writers of fan fiction), took down hundreds of journals in response to concerns about child pornography. After thousands of customers rebelled, LiveJournal restored most of the journals, but it and other sites retain the right to take down content without warning.¹²

Clearly then, as formidable as the challenges associated with archiving and collecting content on media that arrives at an archive as part of an author's collection, the issues for cultural heritage materials that are not immediately accessible from physical storage media are more daunting still.¹³ Given that cloud or network-centric computing will only become more common, it may be, in fact, that content from the first twenty-five years or so of personal computing represents an anomalous window of opportunity where the archivist has reasonable prospects for access to the original storage media.

For all of the challenges and concerns that exist, we would not want to fail to convey the excitement we collectively feel about the unprecedented opportunities for research and scholarship also posed by born-digital literary collections. While there have been numerous studies of the impact of computerization on composition, these often present the computer as simply another tool or instrument.¹⁴ In fact, we believe a computer functions much more like an *environment*—or a writing space, to use a term popularized by Jay David Bolter. Access to an entire computer is not unlike having a key to an author's study or workroom. On the one hand, this creates an awesome burden of responsibility for the archivist, since all manner of sensitive personal information can be inappropriately exposed. Sometimes donors will make their wishes explicit in this regard; but what if they don't? Should a researcher be allowed to see an author's choice of desktop wallpaper? After all, scholars have traditionally been interested in the physical setting in which an author worked, even to the level of such details as furnishings, decorations, and, yes, wallpaper. Other examples begin to enter into the realm of forensic information recovery. A computer's registry, for example, stores information related to all of the device drivers and application software in the operating system. Access to the registry is among the most invasive procedures an outsider could undertake; but its value as a record of the digital

¹¹ <http://news.zdnet.co.uk/internet/0,1000000097,39195962,00.htm>.

¹² http://news.cnet.com/Mass-deletion-sparks-LiveJournal-revolt/2100-1025_3-6187619.html.

¹³ The most thorough study to date is Simson Garfinkel and David Cox, "Finding and Archiving the Internet Footprint," paper presented at the First Digital Lives Conference, British Library, February 9-11, 2009.

¹⁴ See, for example, Daniel Chandler, *The Act of Writing: A Media Theory Approach* (Aberystwyth: University of Wales, 1995).



environment of the computer is enormous. At the level of individual works, scholars will surely want to examine a file's properties, which contain records of when it was last opened and closed and how many hours and minutes was spent accessing it. This kind of metadata, while hardly infallible—it could be spoofed by something as simple as an incorrect system clock—could, with care, be used to establish chronologies that could date the composition of a work—or specific passages within a work—to the hour, minute, and second.

Likewise, various word processing packages incorporate “track changes” features which preserve a record of a document's internal edits, as well as marginal commentary. “Track changes” is already a widely used editorial tool (see our example of Tom Zigal, above), and can systematically capture the kind of revision history for a document that had heretofore been available only incidentally. Given the ease with which multiple versions and drafts can be saved—often this occurs automatically as a function of the software—it is not hard to imagine a scenario in which a scholar may potentially have access to hundreds or even thousands of versions of the same work, and be faced with the prospect of discovering what significant differences between them actually exist. Here we may see textual scholarship begin to draw heavily on text mining and visualization, methods which are specifically aimed at sorting and sifting large volumes of data. For example, a scholar might use a combination of data mining and visualization to discover “hot spots” in the evolution of a work, points at which especially significant revision activity took place. These would then become the basis for further study.

We believe the need for ongoing and open communication between archivists, authors, and the scholars who will eventually use this material is paramount. Even in the brief discussion above, the roles played by each of these actors is sometimes ambiguous. How much user expertise should a scholar be expected to have when it come to examining born-digital data from legacy systems? Is the ability to run a hex editor on a file recovered from a Commodore-64 the equivalent of the assumption that a scholar interested in the early modern era will have appropriate training in paleography? To what extent does the research space needed to support this scholar's work replicate the forensic workstation of the digital archivist? How can authors make reasonable decisions without the expert consultation of the archivist about what kinds of access to their materials should and should not be permitted? How can archivists best anticipate the needs and interests of scholars? Each of these constituencies has insight and expertise that can benefit the others.

Finally, the need for basic (cyber)infrastructure in the area of born-digital collections is clear. In the short time the project was active, through a series of serendipitous encounters, we doubled or tripled the pool of potential collaborators for future efforts. Infrastructure could and probably should take the form not only of enhanced communications, but also provisions for shared access to hardware. It is unrealistic to expect any one institution to maintain functioning instances of every conceivable type of platform that might eventually be required to access data from an obsolescent media format; by the same token, it is simply unnecessary for every institution to duplicate the time and the trouble required to maintain even the most commonplace platforms.



Mature cyberinfrastructure will also take the form of tools for collaboration, shared data and metadata, and federated indices.¹⁵

VII. Next Steps

As we hope is evident from the preceding sections, developing approaches to managing and collecting born-digital content is a timely and urgent topic of conversation. A small planning grant with modest aims resulted in numerous connections with other institutions and programs. The Bodleian, the British Library, the Beinecke at Yale, and the New York Public Library are all now potential partners in our efforts going forward. In addition to this white paper, which will itself have a public life on the Web, we have created (as noted above) a Google Group as a stop-gap measure toward attaining basic cyberinfrastructure in this emerging area of effort. This group is administered by Gabriela Redwine and has 16 members at the time of this writing.

In the short-term, contacts made during the grant period resulted in invitations to group members to two events in early 2009: the first was a research conference hosted by the Digital Lives project at the British Library in February 2009; Farr, Nelson, and Redwine all attended and gave presentations. The second was a one-day meeting, “Stewardship of E-Manuscripts: Advancing a Shared Agenda,” just prior to the DigCCurr 2009 conference in Chapel Hill, North Carolina. Farr, Kirschenbaum, and Peters attended and gave short presentations. Kirschenbaum also presented the project at the Modern Language Association (December 2008), the O’Reilly Tools of Change publisher’s conference in February 2009, the Society for Textual Scholarship in March 2009, and at the Beinecke Library in April 2009.

An article in the February 19th, 2009 edition of the *Times Higher Education* reported on the Digital Lives conference and quoted Nelson.¹⁶ Her comments on the opportunities and challenges inherent in born-digital records reflected the joint explorations undertaken during the grant. In April 2009, the *Chronicle of Higher Education* published a front-page story based in large part on the activities funded by the grant; it included multiple quotations from Kirschenbaum, Nelson, and Reside.¹⁷

As noted above, during the Austin meeting Kirschenbaum and Richard Ovenden at the Bodleian

¹⁵ One outcome of the “Stewardship of E-Manuscripts: Advancing a Shared Agenda” meeting held prior to the DigCCurr 2009 conference in Chapel Hill, North Carolina was the formation of a Personal Digital Archive Working Group (“PDAWG”). Contact Cal Lee, Assistant Professor, School of Information and Library Science, for further information: callee@email.unc.edu. Materials related to the meeting are available here: <http://ils.unc.edu/callee/emanuscripts-stewardship/index.html>.

¹⁶ Available here: <http://www.timeshighereducation.co.uk/story.asp?storyCode=405411§ioncode=26>

¹⁷ “Archiving Writers’ Work in an Age of Email,” *Chronicle of Higher Education*, April 10, 2009: A1.



agreed to take responsibility for coordinating funding for several additional symposia and workshops. A proposal for the first of these, as well as an accompanying CLIR report on born-digital archives and computer forensics, is currently pending at a foundation; if approved, the first workshop is tentatively scheduled for May 2010 at the University of Maryland.

Going forward, the group found no shortage of areas in which substantial future collaboration would be beneficial. At the same time, it is clear that we will need to focus our efforts on what we can do best, given our combined interests and unique expertise, and given the nature of our respective collections. In particular, we are keen to capitalize on our strength in combining archival and information science with scholarly interests. We therefore envision a broad focus on use cases and access strategies for our future work, with different institutions cultivating and deploying approaches depending on local needs and circumstances. Emulation is likely to form a cornerstone for this work, since it currently offers the most promise of retaining the “materiality” of the original digital environment. We are actively researching appropriate programs at the NEH, the IMLS, and other agencies for sponsoring this work.

Likewise, we anticipate continuing to build connections to scholars and other users on the one hand, and are looking forward to working with more authors on the other. Contact among these groups can be difficult to achieve because their professional boundaries do not always encourage interaction, but based on both the internal dynamic within our own group, as well as our conversations with writers during the course of the grant, we believe that more such interaction will be invaluable. Moreover, because some electronic-only literature imagines the reader as co-creator (*Marble Springs* is a case in point), it may also be worth asking what kinds of research environments, collecting policies, and curatorial approaches might be developed to support and preserve user-generated content.



Appendix 1. Meeting Agendas

Maryland Project Meeting Agenda

Friday, September 26th

Location: MITH Conference Room (basement level of McKeldin Library)

Attendees: Rachel Donahue (UMD), Erika Farr (Emory), Matt Kirschenbaum (UMD), Kari Kraus (UMD), Naomi Nelson (Emory), Catherine Stollar Peters (NY State Archives), Gabriela Redwine (Texas), Doug Reside (UMD)

9:30 - 9:45 Introductions, Welcome from Neil Fraistat (Director of MITH)

9:45 - 10:30 Project Overview: Preliminaries and Objectives

10:30 - 12:00 Discussion and Examination of Deena Larsen Collection (Matt)

12:00 - 1:30 Lunch w/ Local Creative Writers (Elizabeth Arnold, William Loizeaux)

1:30 - 2:00 Related Work I: Jonathan Larson's Digital Papers (Doug)

2:00 - 2:30 Related Work II: PARADIGM and The Variable Media Network (Matt)

2:30 - 3:00 Related Work III: Preserving Virtual Worlds (Kari)

3:00 - 4:00 Summing Up: Common Interests, Common Concerns

4:00 - 4:30 Project Infrastructure: What's Needed at this Stage?

4:30 - 5:00 Agenda Planning for Austin Meeting (Gabby, Chris Grogan)



Schedule for Harry Ransom Center site visit, 12 Nov. 2008

Participants

Emory: Erika Farr and Naomi Nelson

MITH: Matt Kirschenbaum, Kari Kraus, Doug Reside, Neil Fraistat

HRC: Catherine Stollar Peters and Gabriela Redwine

Yale: Michael Forstrom and Tom Hyry

Location unless otherwise noted: Denius Room, 2nd floor

- 8:45 Arrive at reading room, check bags, and get settled.
- 9 - 9:30 Welcome from Dr. Staley, introductions, project overview and review
- 9:30 - 10:15 Tour of digital processing area (4th floor)
- 10:15 - 11:30 Michael Joyce collection
- 11:30 - 12:30 Pat Galloway, current digital preservation work at the iSchool
- 12:30 - 1:30 Informal lunch with Zach Vowell, archivist of the Video Game Archive (Moseley Room, 3rd floor)
- 1:30 - 2:30 Arnold Wesker collection
- 2:30 - 3 Carlos Ovalle: intellectual property, copyright, digital materials
- 3:15 - 4:15 Summing up (Post-Its)
- 4:15 - 4:45 Agenda Planning, including for next day's Ovenden meeting (*We'll need to leave the Denius room at 4:45 but can continue the conversation in one of the 3rd floor break rooms if we need extra time.)



Itinerary for the Emory Site Visit, December 4th through December 6th:

Thursday, December 4th:

Afternoon: Project partners arrive at Emory Conference Center in Atlanta, GA

Useful information:

Hartsfield-Jackson Airport: <http://www.atlanta-airport.com/>

Emory Conference Center info: <http://www.emoryconferencecenter.com/>

Emory campus map: <http://map.emory.edu/>

Friday, December 5th:

Breakfast on your own.

Unless otherwise noted, meeting events occur in the Kennesaw Conference Room (Room 225 in the North Decatur Building)

9:00am – 9:15am Welcome (Rick Luce and Joan Smith)

9:15am – 9:45am Introductions and Updates

9:45am – 10:30am Rushdie's Computers

10:30am – 10:45am Break

10:45am – 11:15am Other Born Digital collections in MARBL (Naomi Nelson, Laura Carroll)

11:15am – 11:45am Digital Curation Center support for born digital initiatives at Emory (Lars Meyer)



12:00pm – 1:30pm (Woodruff Room, 10th floor, Woodruff Library) Lunch with Natasha Trethewey. Discussion topics to include literary processes, understanding digital media in relation to literary “papers,” how researchers might engage with such media

1:45pm to 2:15pm MetaArchive and Digital Preservation Initiatives (Katherine Skinner)

2:15pm to 3:15pm Presentation by and discussion with Bill Underwood on processing born digital records from President Bush

3:15pm to 3:30pm Break

3:30 to 4:45pm Grant Administration (Review outcome and discussion topics from previous meetings, Post it Note exercise, identify themes and conclusions from meeting, etc.)

4:45pm to 5:00pm Wrap-up (Identify next steps and sketch agenda and outcomes for Saturday meeting)

Saturday, December 6th:

Breakfast on your own.

Location: ECIT 214, Woodruff Library, 2nd floor

9:00am to noon: Follow up from day one, begin collaboratively drafting grant documents, create a plan for grant outcomes

Afternoon: Project partners head back home.



Appendix 3. Participant Biographies

Erika Farr is the Director of Born-Digital Initiatives at Emory University’s Robert W. Woodruff Library. She has extensive library experience with digital text production, digital preservation, and subject-driven portal development as well as a doctorate in English Literature that provided her with training in traditional research methodologies. Her current research includes digital curation, and, in particular, the scholarly implications of processing and providing access to born-digital and hybrid archives.

Matthew G. Kirschenbaum (Project Director) is Associate Professor in the Department of English at the University of Maryland and Associate Director of the Maryland Institute for Technology in the Humanities (MITH). He is also an affiliated faculty member with the Human-Computer Interaction Lab at Maryland, and a Vice President of the Electronic Literature Organization. He has participated in sponsored research on text mining and visualization, as well as the preservation of interactive virtual worlds. Kirschenbaum’s first book, *Mechanisms: New Media and the Forensic Imagination*, was published by the MIT Press in 2008. He has published and lectured widely in topics in digital humanities ranging from electronic literature to interface and visualization to text encoding and textual editing.

Kari M. Kraus is an Assistant Professor in the College of Information Studies and Department of English at the University of Maryland. Her research and teaching interests focus on new media and the digital humanities, digital preservation, intellectual property, virtual worlds, and textual scholarship and print culture. Kraus is a co-principal investigator on the Preserving Virtual Worlds project sponsored by the Library of Congress and the principal investigator for the Digital Humanities Model Internship Program funded by the Institute of Museum and Library Services. She has taught at the University of Rochester and the Eastman School of Music, and in the Art and Visual Technology Program at George Mason University.

Naomi L. Nelson is the Interim Director for the Manuscript, Archives, and Rare Book Library (MARBL) at Emory University. She received a Masters in Library Science from the University of Pittsburgh in 1991 and a Ph.D. in History from Emory University in 2001. She has been working with electronic records for over ten years, starting in 1996 when she served as a consultant to the Senate Computer Center on the transfer of born-digital Constituent Mail index files to archival repositories. She has taught workshops for the Society of American Archivists (SAA) on the “Digitization of Archival Materials” and “Digital Libraries and Digital Archives.” She was a member of the Collection Working Group of the Digital Libraries Federation’s Aquifer project, and currently chairs the SAA Technology Best Practices Task Force and serves a consultant to the MetaArchive Project funded by the Library of Congress’ National Digital Information Infrastructure and Preservation Program.

Catherine Stollar Peters is an Archives and Records Management Specialist for the New York



State Archives. In her previous position as Archivist at the Harry Ransom Center at the University of Texas at Austin she specialized in electronic records preservation. At the Ransom Center, she developed policies and procedures for archiving electronic records. Her work with the archival materials of hypertext author Michael Joyce is the subject of multiple publications including a case study recently published by the Society of American Archivists in the symposium proceedings of *New Skills for a Digital Era*. Currently, she is working on her Ph.D. in Information Science at the University at Albany with a focus in archiving digital manuscripts and archival forensics.

Gabriela Redwine is an archivist and electronic records specialist at the Harry Ransom Center, where she is responsible for developing and implementing digital preservation policies and procedures. She earned her B.A. in English from Yale University and her M.S. in Information Studies and M.A. in Women's and Gender Studies from the University of Texas at Austin.

Doug Reside is the Assistant Director of the Maryland Institute of Technology in the Humanities (MITH) and a Visiting Assistant Professor of Theater at the University of Maryland in College Park. Doug holds undergraduate degrees in Computer Science and English and earned his Ph.D. in English at the University of Kentucky where he worked on several digital humanities projects, including Kevin Kiernan's celebrated *Electronic Boethius*. Reside's primary research interest is musical theater and the way in which digital technology can be used both to create and to preserve the art form. In addition to his managerial, and programming work at MITH, Doug is currently working on a book on the "born-digital" musical.



Figure 5. Reside, Farr, Kirschenbaum, Kraus, Nelson, Redwine, Peters