Title
Designing a Born-Digital Archive

Permalink
https://escholarship.org/uc/item/8wf5w4nk

Author
Light, Michelle

Publication Date
2010-05-14

Presented at "Time Will Tell, But Epistemology Won't: In Memory of Richard Rorty"

A Celebration of Richard Rorty's Archive

May 14, 2010

University of California, Irvine

Online finding aid for the Richard Rorty papers: http://www.oac.cdlib.org/findaid/ark:/13030/kt9p3038mq/

Richard Rorty Born-Digital Files: http://ucispace.lib.uci.edu/handle/10575/7

Archivists often don't get much recognition when they process a collection and make it available for researchers, so you can imagine my honor at being here today, knowing that our hard work inspired a conference with such wide ranging scholarship.

So first, I would like to recognize the library staff and our volunteers who made this project possible.

First, our temporary archivist, Dawn Schmitz was responsible for processing the Richard Rorty papers. She transformed a chaotic mess into a beautifully organized, preserved, and described collection. The Richard Rorty papers span 61 boxes of archival materials (the equivalent of approximately 50,000 pages) and include over 1,000 digital files. Three UCI graduate students Ali M. Meghdadi, Brian Garcia, Tae-Kyung Timothy Elijah Sung, helped Dawn organize the materials.

In the course of processing, Dawn discovered over 70 floppy disks. A cursory review of the material revealed that the content on the disks largely did not duplicate the paper records, and there was a substantial amount of content on those disks. Fortunately this discovery coincided with my work on launching the UC Irvine Libraries' Digital Scholarship Service. With a team of librarians, we were setting up a platform to preserve and provide access to UCI's unique digital research assets. The born digital materials from Rorty provided the perfect pilot project for us to build and test the capabilities of our new service, UCIspace @ the Libraries. So I would also like to recognize Caryn Neiswender, who was pivotal in setting up UCIspace @ the Libraries; Holly Tomren, our metadata expert, and web designers, Mark Vega and Sylvia Irving.

By born-digital, I am referring to records that were created digitally and usually only exist digitally. I am not talking about digitizing, or scanning paper-based documents. In the born-digital collection of Richard Rorty in UCIspace @ the Libraries, we only have in there the files that Rorty created and maintained digitally. All his paper files are open for researchers to use in our Reading Room still. The Richard Rorty papers are truly a hybrid collection. To understand his work, you need to spend time in both his paper and electronic files.

Fending Off Digital Decay, Bit by Bit
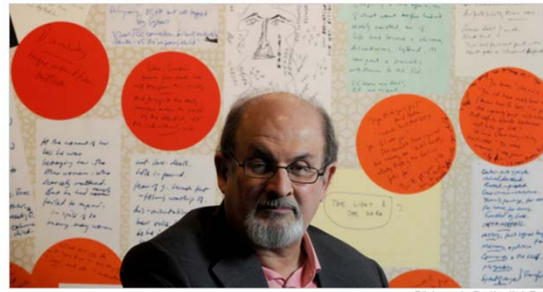
Salman Rushdie at Emory University in Atlanta, which is currently exhibiting his personal archive, including personal papers, and electronically produced drafts of his novels.

By PATRICIA COHEN
Published: March 15, 2010

Among the archival material from Salman Rushdie currently on display at Emory University in Atlanta are inked book covers, handwritten journals and four Apple computers (one ruined by a spilled Coke). The 18 gigabytes of data they contain seemed to promise future biographers and literary scholars a digital wonderland: comprehensive, organized and searchable files, quickly accessible with a few clicks.

In March, *The New York Time*s featured an article called, "Fending Off Digital Decay, Bit by Bit," which aptly summarized an issue that keeps me up late at night: how can the UCI Libraries preserve archives that are increasingly created and only maintained in digital form. For decades, we've managed to preserve hundreds of linear feet of records printed on paper, and I'm confident that in a hundred years, scholars will still be able to read and study this work, but I definitely can't say that with confidence about the archives created in the near past, today, and in the future in digital form. There are many questions:

- Will we have the right hardware or software to read the sometimes mysterious digital media we find in boxes 15 years after they were created?
- Even if we found the right software and hardware, will the bits…the 0s and 1s that carry the content…still be there for us to read?
- How much do we need to preserve to support the scholarship of the future? The content only? Or perhaps the context of author's computing environment as well?
- How can we be sure that no one has tampered with the files over time, that we are presenting users with the files as the author originally intended?
- How widely should we distribute, e.g., on the web, content that authors and their subjects never intended to be made public? How much should we balance respect for privacy of those who are alive with scholars' need to know?
- Are authors saving their files these days in ways that will support the archive and scholarship of the future? How will a computer-generated text transform or hamper the work of scholars who study the evolution of a text over time?
- And the kicker, particularly in this new budget climate, will we have the resources, expertise, or support within UCI's mission to do this kind of work in the future?

I recommend this *New York Times* article for those just learning that archiving born digital materials is a challenge, as it describes some interesting approaches that Emory University is using for the Salman Rushdie Archive or how Stanford University is creating a digital forensics laboratory to recover data. But the Harvard approach is far more common in the field. The curator at Harvard explained that, "We don't really have any methodology as of yet" to process born-digital material. "We just store the disks in our climate-controlled stacks, and we're hoping for some kind of universal Harvard guidelines," she added.

Here, we don't have the budget of Harvard, Stanford, or Emory, but we also have born digital materials that to which people wanted access. In response to numerous people showing interest in writing dissertations based on the Rorty papers, and because we knew that floppy drives were increasingly rare on campus and that the disks' life spans are relatively short, we had to act now or risk forever losing their contents. We didn't wait for clear best practices or standardized methods from the profession. We simply did our best to preserve and provide access to the materials with our available resources.

Technological Obsolescence and Frailty

What does it take to preserve a born digital archive?

When encountering born digital materials in archival collections, three of the questions I ask first….
- Do we have the right hardware to open the files, for example, if they are on zip disk, tapes, or floppy disks?
- Do we have the right software or operating system to open the files?
- Have the bits on the media survived?
  - Bits can rot or decay. Bit rot happens when the small electric charge of a bit in memory disperses, possibly altering code.

Digital media is fragile; its shelf life compared with, say, non-acidic paper is extremely short. Perhaps even more intractable, is the rate at which computer hardware and software become obsolete.

For the Richard papers, we were fortunate enough that three things coincided: 1) we still had access to computers with floppy drives, 2) Rorty mostly used word processing programs that created files that Microsoft Word could still open, and 3) very few files were corrupted. In order to ensure that scholars of the future would also be able to open to files, we decided to transfer all of his files into PDF/Archive files, because the PDF/A format is intended to be a long-term preservation standard.

But for other collections we haven't been so fortunate. In some other collections, I've encountered 8 inch floppies, 5 ¼ inch floppies, and Apple zip disks for which we can't find the computers on campus to open them. The reality is that most archival repositories, including our own, don't have the staff, space or resources to become a technological museum with many kinds of equipment, operating systems, and software packages. As an archivist, I faced long ago that I don't have the resources to save everything, and hard choices have to be made daily about what to carry forward with us from the past.
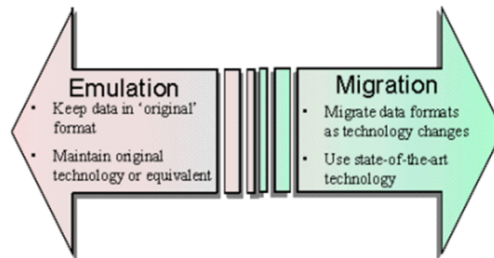

When we made the decision to transfer Rorty's files from their original file format to PDF/A, we performed "migration." That is, we migrated the files from an obsolete platform to another. We knew there were some trade-offs, which we thought were defensible given the nature of the content, that is, predominately word processing files that were meant to be read as stand-alone objects.
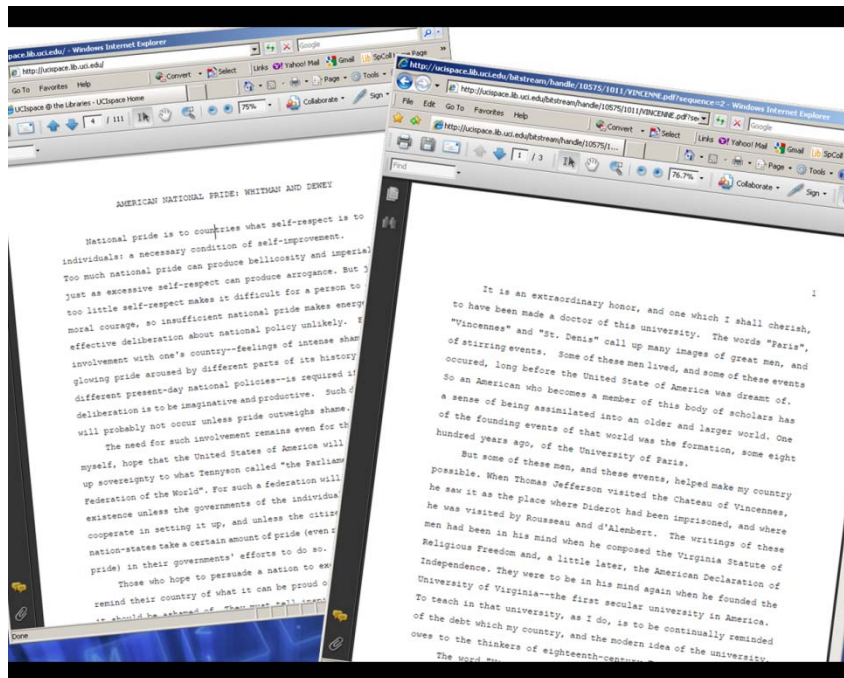
Nevertheless, we lost the original functionality and the look and feel of the programs Rorty used. There may have been some loss or addition of information, which Dawn will address later. And users will have to take it on faith that the documents are pretty much the same as the original. It's also hard to get a feel for what the disks were like, or how Rorty used them to save information. In other words, we opted to preserve content but lose some context. We would have never been able to complete the project if we tried to emulate Rorty's original computing environment for you.
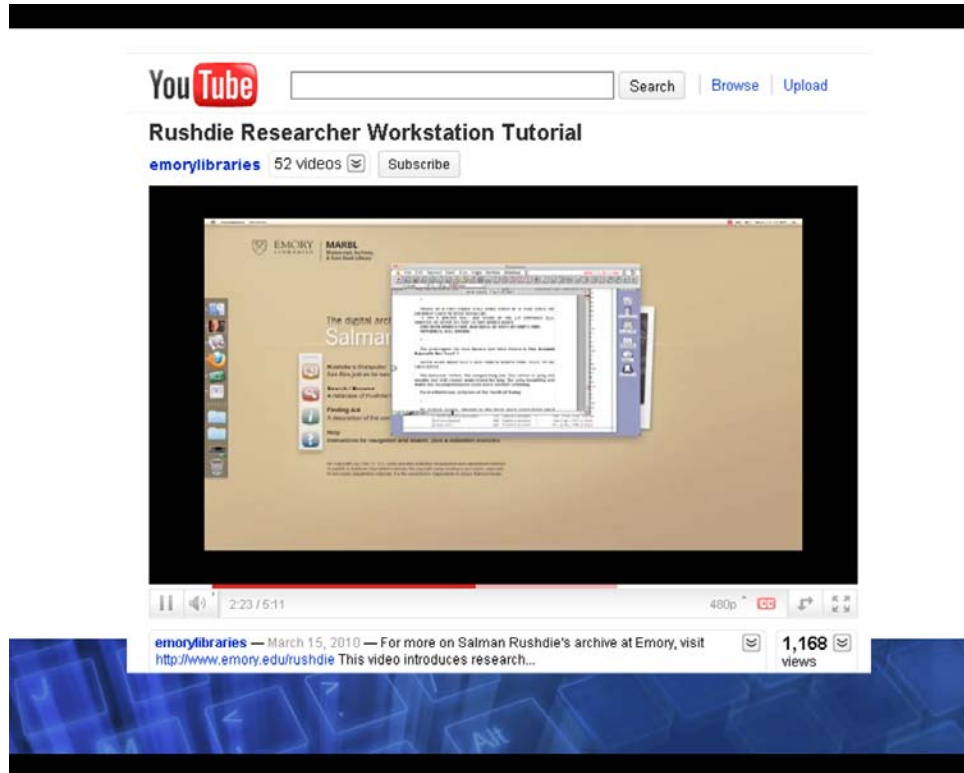
Emulation is the competing option to migration, and it essentially involves emulating the original computing environment to preserve its look and feel and interactivity. Emory University, for example, has emulated Salman Rushdie's Macintosh, so if you visit Emory you can see how Rushdie organized his files, or you can open his calendars, relive his experience in writing a novel in the software and operating system he used at one time, or see how he used virtual sticky notes to record notes. The director there believes in preserving the whole ecosystem, or "biostructure," of the author's digital archive: the hardware, software, programs, and applications, all the files and file names, search histories—even the order in which everything was installed. "There is something fundamentally interesting about the computers themselves," she says, "as the medium between the user and the digital media." For those who won't get to Emory to experience this, you can watch a You-Tube video of how the system works.

Emulation is extremely expensive and difficult, and I don't see us trying to do this in the future. However, not all projects will be as simple as the Rorty papers. As we must preserve more contemporary ways of authoring content, for example, if authors are using audio or visual content, or are relying on hyperlinks to draw together many files, using a variety of email accounts, using social networking sites, or having conversations that span multiple websites, we will probably have to investigate preservation of this "biostructure." We increasingly need to preserve context when it is often just as important, if not inseparable, from the content.

Authenticity and Integrity

Archivists are very concerned about the authenticity and integrity of the records they are entrusted to protect. In fact, when you come into an archive, I'm guessing that most of you trust that when we give you the papers of an author, you can believe that most materials are truly created or collected by that person and that the materials are unchanged from how the creator intended them to be. There might be various clues that they are not, such as someone else's handwriting, to signal to you if someone else made changes to a document or if someone else's materials got mixed up in a folder. Or you might notice when a part is missing or has been defaced.

As you know, digital content is highly mutable. It is easy to make changes and wipe away evidence of the past. Archivists are finding that they have to come up with new procedures and systems to make the same sort of assurances about the authenticity and integrity of someone's electronic files, so you can trust that the files you see are the ones the author created.
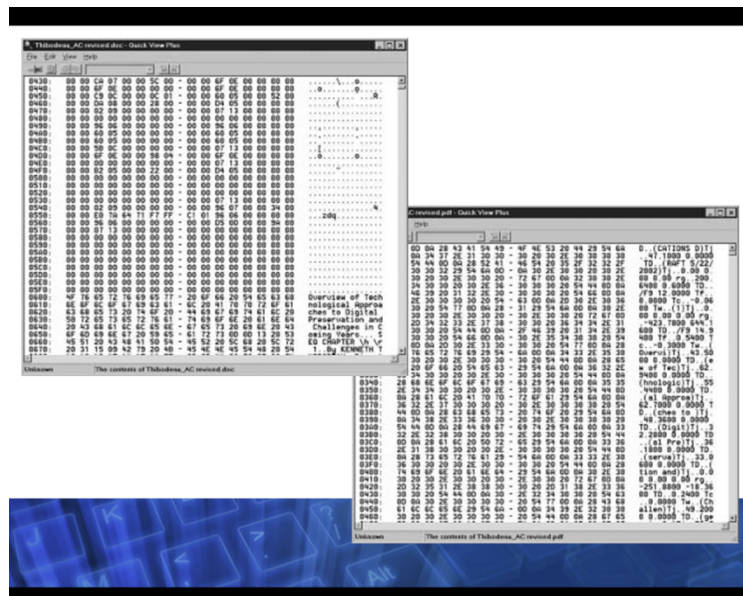
We might also ask, what is the record we are preserving anyway? On this slide, you'll see the same paper, one in Word and the other in PDF. Each displays the hexadecimal values of the bytes that encode the beginning of the document. Neither is the exact equivalent of the conceptual document. Both contain the title of the article, but otherwise they differ substantially.

When we migrated Richard Rorty's files, we transformed the underlying code, that is, we tampered with the integrity of each digital object, even though we've strived to present an unaltered conceptual record.

After we put material out on the web, we also have to be careful that the files don't get corrupted. We are taking security and preservation precautions so that you may continue to believe that you are looking at Richard Rorty's work.
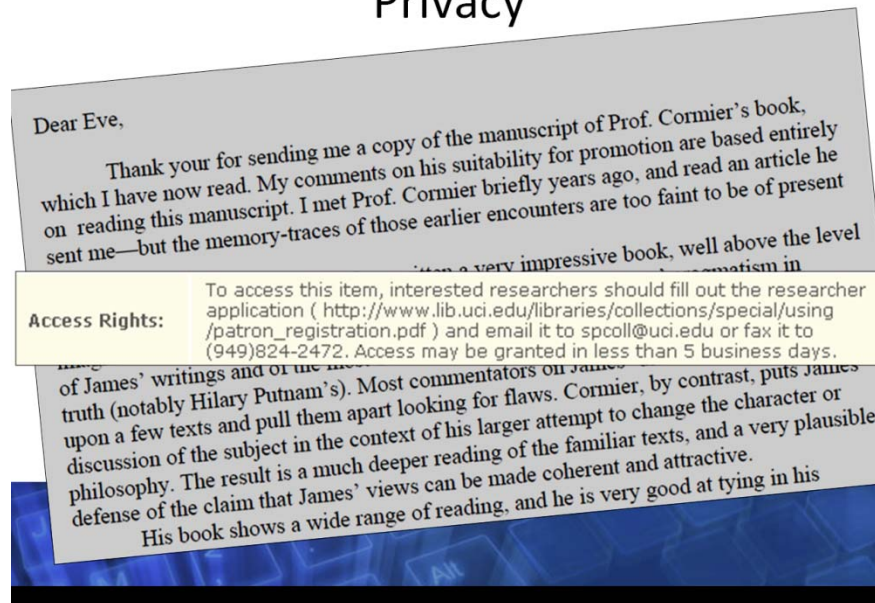

Copyright

From the Library of Congress, "Taking the Mystery out of Copyright"
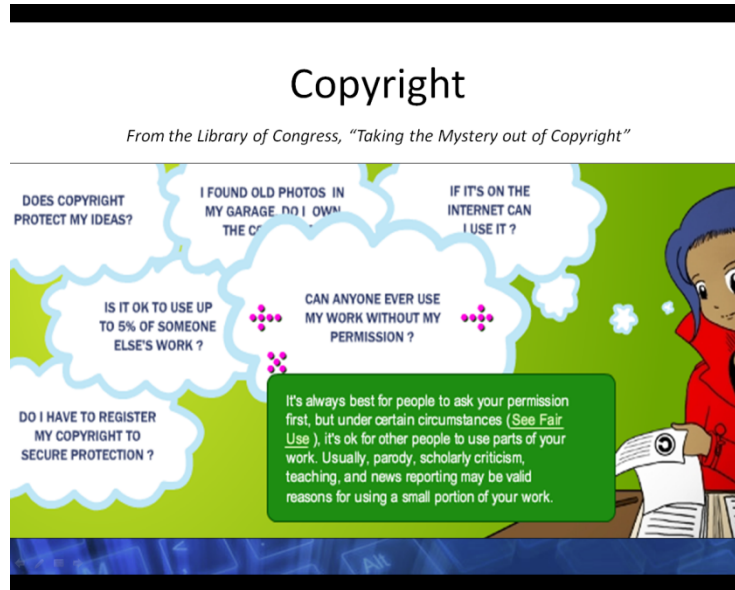
It is standard for Special Collections to allow the donors and their heirs to retain control over their literary rights or copyrights, even if though they give us physical ownership over their materials. Rorty and his heirs retained these rights. The gift agreement said very specifically that the Rorty heirs, not the Libraries, retained publishing rights in particular. This is one of a few reasons for why we created a virtual reading room in UCIspace @ the Libraries. When he gave UCI his papers, Rorty also stated we should provide access to his materials according to our standard procedures. Our standard procedures are essentially that researchers fill out an application to use materials in our reading room and agree to follow our rules, one of which is that they respect copyrights. The virtual reading room allows us to preserve the material on the floppy disks and provide scholarly access to them in their original digital form, but we're not infringing on the Rorty family's right to control their formal publishing and wide distribution. For example, Rorty's heirs still retain the right to grant permission to someone to create and publish a critical edition.

We've also discovered that even though other institutions also have virtual reading rooms, you still have to go to the institution to use the materials online. We are pushing the envelope a bit by allowing scholars to access the virtual reading room remotely, but we believe we are fairly balancing copyrights, the terms of our gift agreement, and access.

For those who don't know, for unpublished materials, authors and their heirs retain copyright for 70 years after the death of the author. Libraries may make copies for preservation purposes or we may make a limited amount of copies for researchers for personal or research purposes without permission, but both libraries and researchers must ask permission from the copyright owner if we wish to reproduce the work beyond what is defined as "fair use" and to distribute it widely. This has stalled many digitization projects in their tracks, especially for correspondence, because it takes time to seek permission from every author for the right to put their letter up on the web.

Privacy

Dear Eve,

Thank your for sending me a copy of the manuscript of Prof. Cormier's book, which I have now read. My comments on his suitability for promotion are based entirely on reading this manuscript. I met Prof. Cormier briefly years ago, and read an article he sent me—but the memory-traces of those earlier encounters are too faint to be of present

...tten a very impressive book, well above the level ...matism in

Access Rights: To access this item, interested researchers should fill out the researcher application (http://www.lib.uci.edu/libraries/collections/special/using /patron_registration.pdf) and email it to spcoll@uci.edu or fax it to (949)824-2472. Access may be granted in less than 5 business days.

of James' writings and of the ... truth (notably Hilary Putnam's). Most commentators on James upon a few texts and pull them apart looking for flaws. Cormier, by contrast, puts James discussion of the subject in the context of his larger attempt to change the character or philosophy. The result is a much deeper reading of the familiar texts, and a very plausible defense of the claim that James' views can be made coherent and attractive. His book shows a wide range of reading, and he is very good at tying in his

In addition to copyright issues, another argument for the the virtual reading room is to respect privacy. We might ask, how much online access should we provide to private, unpublished materials? Should they be indexed for the whole world to discover in Google? What should guide our decisions? Privacy laws? Fear of lawsuits? Sensitivity to the privacy of our donors? Sensitivity to the privacy of individuals that may be mentioned in the papers? Time?

As an academic, Richard Rorty was asked to write countless recommendations of students and colleagues for admission to graduate school, for job applications, or the tenure process. Federal law protects the privacy of student records no matter where they are found, so we knew that we needed to remove and protect these for the lifetime of each student. But what about those recommendations for former students or colleagues? These aren't exactly protected by law. If Rorty wrote you a letter of recommendation, would you want us to put it up on the web for Google to index it?

We decided that we wanted Rorty scholars to have access to these, because there is a lot of significant content in there revealing Rorty's thought, but we didn't want to open these up for indexing in Google for access by unknown individuals with unknown purposes. We decided it was safest to provide access to these materials in the same way as we always have, in a reading room that scholars come to and who agree to use the content according to some basic rules. We still, however, are widely distributing the metadata about Rorty's files so interested scholars can find us. We thought this way we would still be faithful to the trust Rorty placed in us to provide access to the materials in the same way we do for other critical theorist's papers in our possession, or for that matter, the trust he shared with his students and colleagues for whom he wrote these letters.

The work of the archivist
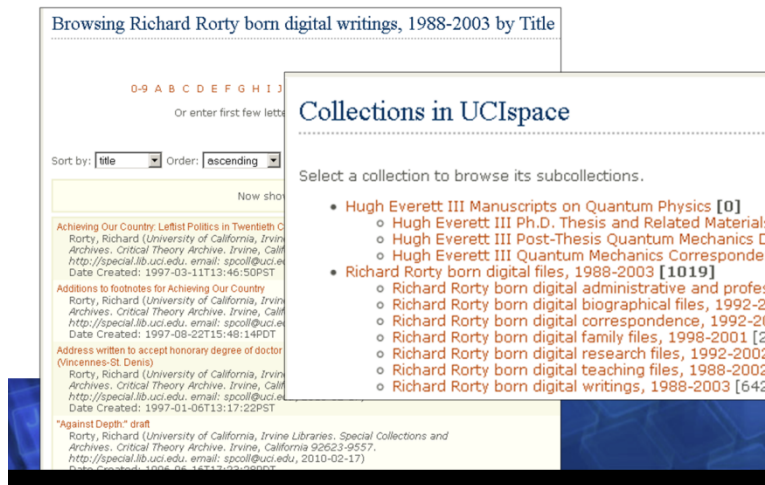
You might think that digital records might make our jobs easier as archivists, but behind the curtain, we still have work to do. In the paper world, you come into an archive, and you may see many items in folders. To make our work more efficient, we try not to handle or describe every single item. However, for this project, we discovered that Dawn had to open up every single file and gave it a title. We could have automated the process, and just used Rorty's eight character file names or used the first 100 characters of each file, but we thought it would be more useful for you to have carefully crafted titles that reflected the content of each file, or so that you could browse. True, the content is all searchable, but you still need to know whether you should click on a search result or not. We also spent time organizing the digital files, to match the organization of the paper files and make it easier to discern teaching materials from the draft manuscripts, for example.

My point here is that archivists' decisions and practices will continue to influence how you may encounter, experience, use, and understand the archive. As archivists, we are mediators between you and texts; we have always had a transformative impact on an archive before you get to it in a repository.

The functionality we build into access systems may also really impact how you do scholarship. Full-text searching is an obvious boon. What if in the future, we gave you the ability to automatically compare drafts to see differences? Or if we allowed you to annotate files and share your interpretations with other researchers?

Managing your digital files for the future

- Use common file formats and software.
- Keep up to date.
- Organize and name files appropriately. Make sure they are self-documenting.
- Back-up your files.
- Protect your files.
- Be aware of privacy and copyright issues.
- Manage your emails. Delete what is not important.
- Save drafts of significant works to document major changes.

It's not always obvious early in a person's career that their materials belong at an archive. Usually we come in late in a person's career. If someone has been creating digital files for 40 years, who knows what may survive? So I will now conclude with what you can do with your own digital files to help preserve your own archive, particularly if one day, an institution may be responsible for providing long term access to your work.

First, manage your own files for the future. These are all common sense tips.

- Use file formats and storage mechanisms that last. Don't use uncommon or propriety file formats that can't easily be opened up by other programs. When necessary, convert outdated file formats to the latest file form
- Back up your data often and keep the back up elsewhere. Make sure you're not storing important files on outmoded storage devices.
- Practice good naming conventions and organize your files logically so you remember what the files are and so that others can easily discern what files might be. You can even add metadata to describe your most important files within the file itself. Most programs have a "Properties" feature where you can add descriptive information about your file.
- Protect your data so it can't easily be tampered with. Use virus software, passwords, firewalls, etc.
- Be aware of privacy and copyrights. Segregate confidential information.
- Manage your email. Only retain significant email or significant files. Even though we can search your content, if you retain a lot of insignificant information, the gems are harder to find.
- Retain drafts of significant works in a logical way

In many author's papers…physical papers….you often will encounter a number of drafts. Drafts often reveal the development of an author's thoughts, but with the ease of the "save" feature in modern word processors, I worry that we may lose the ability to study this evolution of thought in the future. It's ironic that "save" can cause so much destruction. So occasionally save drafts of significant works, so you may preserve important changes of thought.

UCI Libraries' Digital Scholarship Service

The UCI Libraries Digital Scholarship Service helps the UC Irvine community publish, share and preserve the digital products of research. Subject librarians will assess your needs, provide guidance on your rights, and make referrals to appropriate digital services. In some cases, the Libraries may collaborate with you to create customized solutions.

http://www.lib.uci.edu/services/digital-scholarship/

Second, you can publish your materials online for long term preservation throughout your career. In February 2010, the UCI Libraries launched the Digital Scholarship Services. The service is designed to help UCI's faculty and research staff find a place to preserve and publish the products of their research to perpetuate scholarly communication. Subject specialist librarians can help you identify a solution for depositing your research somewhere where it may be available long term.

One common solution is to use eScholarship. With eScholarship, UC faculty, departments, and centers can publish books, articles, journals, conference proceedings (such as this one), and seminars online.

You've all had a chance to use UCIspace @ the Libraries with the Rorty papers. If eScholarship doesn't fit your needs, UCIspace @ the Libraries may be an appropriate home for some of your research materials or any of the other tools we link to from our Digital Scholarship Services page.

For those at other universities, I bet your library also offers a similar service.

In sum, don't wait until it's too late to take action to preserve your born digital materials. The future of this kind of scholarship depends on your good management of your files, you sharing your digital content in places that will be around long term, and your support for your Libraries' services.