The present transcript, based on a video recording of a seminar offered at the 2010 RBMS Preconference, was commissioned in 2013 by the accessibility office at San Jose State University. No attempt has been made to fix errors of transcription.

Original video:
http://www.rbms.info/conferences/preconfdocs/2010/Talks/Seminars/SeminarI.wmv

---

I. Born-Digital Manuscripts: A Primer [video]

Jennifer Schaffner, OCLC RLG Programs (moderator)
Laura Carroll, Emory University [slides]
Erika Farr, Emory University
Michael Olson, Stanford University [slides]
Ben Goldman, University of Wyoming [slides]

Don't be frightened! "Hybrid collections" are a relatively new thing. What are the four or five simple things that everyone – librarian, archivist, administrator - needs to know? The authorities in this seminar will demystify collecting, preserving, describing and providing access to born-digital materials. It's early on, but tools and best practices for holistic management of born-digital manuscripts are emerging. Speakers will share tips, hints and lessons learned from both mainstream and high-profile digital collections.

>> Good morning, hello, hello. We're a friendly bunch, but come on. Take a seat. So good morning. Welcome to the Dee mystification session. This is seminar I, Born Digital Manuscripts, a primer. And I'm Jennifer [Inaudible] C LC research. And collaboration, if anything requires collaboration, it's going to be facing the challenges of hybrid collections. And to give you a preview of Jackie's results of the survey of special collections is coming out in July, which is the way she puts it. Born Digital archival materials, under collected, under counted, under managed, unpreserved, inaccessible. So now we know where we stand. And I'm serious. And she's got the evidence to back that up. So we're living in the first wave of Born Digital materials, that's how I think about it, the first wave. These are Born Digital manuscripts and archives that are in special collections. And I think of these things as artifacts of our near past. And they're a little bit different for us, they're not the digitized materials that we've been dealing with for years. And of course they're at risk. So what are -- this is the question, what are the four or five essential steps that everyone, the librarians, archivists, and especially our administrators, everyone needs to know? It's early on, but tools and best practices for holistic management of Born Digital manuscripts are emerging, but our panelists today are pioneers in the wild west. They didn't wait for clear standards and procedures, they just jumped in. So we're going to hear from Laura Carol and Erika Far from Emory, Michael Olson from Stanford, and Ben Goldman from the University of Wyoming. I'm going to introduce everybody and then let them talk. We're going to aim to speak all together for an hour, and we want a fair, frank Q and A. Erika Far is the director of Born Digital initiatives to emery. And in 2006 she joined a team of archivists, technologists, and libraries working on the literary papers of Salman Rushdie,

which is a hybrid archive of both paper and Born Digital papers. And Erika is currently, believe it or not, finishing up her MLIS, but in 2004 she took her Ph.D. in English from emery, and we bonded because it's on 17th century English poets. And Erika is one of our -- the book person and the scholar on the panel. So Laura Carol is a manuscript archivist at emery where she is the lead on the Salman Rushdie Born Digital archives project. And prior to emery, Laura was an archivist at the American Medical Association, Newberry Library, and Loyola University. She holds a master's degree in public history from Loyola, in addition to an MLIS from Dominican. And she's probably our archivist on the panel. Michael Olson is the manager of digital projects for Stanford University. And Michael works with library curators and other non-library constituents to build and deliver digital collections with the primary focus on unique, special collections. Michael's process, many of Stanford's Born Digital collections, such as Robert Sealy's computer files, and did you have a hand in the files of Peter Koch, I think? Yes, and so before Stanford, Michael took an MFIL in history and computing from the University of Glasgow, and he's got a BA from the University of British Columbia in mediaeval studies. And so Michael is our historian and digital librarian. Ben gold man. As a digital programs archivist at the American Heritage Center in Wyoming, Ben Goldman manages the center's developing electronic records program, as well as it's ongoing digitization activities and the center's web sites. And before becoming an archivist at Syracuse, he worked for several years in corporate IT, most of it as a web designer and administrator, and he wrote the definitive lit review and paper on digital humanities and Born Digital manuscripts. And somebody in the back should tweet this, you can find it on the web, at bgoldman.info /dh/litreview/html. You know, I could put it on the [Inaudible] first let's welcome both Laura and Erika on the Rushdie manuscripts.

[Applause]

>> Okay, all right, Jennifer, thank you very much, and thank you guys for having us. Erika and I are going to be co-presenting this talk. When we first started talking about presenting at this conference and giving the conference theme, join or die collaboration, we really realized it really did exemplify the why in which we work and the way in which we continue to work throughout the project. So I'm going to first provide you with a little bit of background to create the context for the rest of the presentation, and then Erika is going to follow up with a bit more detail. So in October 2006, emery acquired the papers of Salman Rushdie. This was a culmination of a relationship with emery that had begun when Rushdie was invited to the campus to deliver the Richard Ellman lectures in modern literature, a biannual series in which a distinguished writer or critic visits the campus and delivers three lectures and gives a reading. He also joined the faculty of the English development as a distinguished writer in residence and visits the campus every spring to teach a seminar and deliver several talks and lectures. For those of you who don't know, Rushdie is a critically-acclaimed novelist and international figure. Rushdie's Midnight's Children, published in 1981, was selected twice as booker of the bookers, in honor of their prestigious -- the prestigious booker price, 25 and 40th anniversaries. So he's equally well known and infamous for his international attention that followed the publication of his novel, the Satanic Verses in 1988, most notably the fatwa issued by the Ayatollah Khomeini. So the Born Digital --

[Background noise]

>> -- the Born Digital portion of this collection consists of four computers. One desktop and three laptops, one hard drive containing files from the fifth laptop that Rushdie had originally planned to give us but it not, a zip disc and another disc that turned out to be mostly application files. When our director had visited Rushdie's home to survey the collection he asked whether Rushdie had any computers. Rushdie relied that he had several computers in his closet, and luckily he hadn't disposed of his old ones when he got new ones. So emery created a proposal out lining it's desire to preserve Rushdie's digital files along side with his paper files. He was intrigued by this idea of preserving his digital records. The records on his computer were from a very significant time in his life, obviously. This was when he was hiding during the fatwa. All right, so this is a hybrid collection, as the title hints at. We did not just get Born Digital material. We also received a substantial amount of paper material, approximately 100 linear feet of journals, correspondence, writings, subject files, et cetera. The paper portion of the collection went straight to the arrangement and description unit. And we were able to begin processing in mid-2007, and I led that team. Meanwhile, as soon as the acquisition arrived emery formed a working group charged with the task of assessing the new challenges and issues that were involved in preserving and making the Born Digital material available to researchers in an innovative and sustainable way. So I also want to provide you with a little bit of context for the organization at emery. Within the Emory University Library System the Manuscript Archives and Rare Book Library, or MARBL, as we'll refer to it, it functions as a small library within a library. So we have a director, we have units such as arrangement and description of rare books, research services, all reporting to a director. Digital systems and the Born Digital initiatives program are separate departments, and they fall under the chief technology strategist. So the programmers we're going to talk about, they work in digital systems, but they also work on other digital projects, and the Born Digital initiative program encompasses other projects such as electronic thesis and dissertations. And so we'll talk more about the ramifications of this organization a little bit later. So going to hand it over to Erika.

>> For this discussion of our handling of Rushdie's papers we're going to resist the considerable temptation for those of us who often work with systems and infrastructure of detailing the technical approaches to processing such a hybrid collection. And instead, we're going to consider how the people gathered around the collection itself impacted the work of the group. So not so much this, but instead, take a look at the actual people involved. While the specification -- specifications of emery's repository and the innovative archival processing of the materials themselves are certainly relevant, and will be discussed, the focus today here will be on how the collaboration among the diverse group of professionals at emery led to productive conversations, informed decision-making, and efficient project development. In addition to many technical and archival lessons learned while preparing Rushdie's content for researcher use, we also learned there are important benefits to having a diverse team working on hybrid archives. So as Laura mentioned in our opening, we acquired the Rushdie materials in 2006. The equipment itself arrived in early 2007. This collection marked the first time that our special collections required complete computer environments. So their arrival ushered in a slew of questions, decisions, some anxiety, admittedly. So we get in the shop, what do we do now? One of the earliest choices the emery library made regarding these materials was to create a multidivisional working group which Laura referred to just a minute ago. This team, which we very cleverly named the Rushdie Born Digital Archives working group, which we shortened to BDAR to make it sound a little more threatening, included three members from MARBL,

Naomi Nelson, who is a -- at that time our interim director. Susan McDonald's, who is our head of arrangement description, and Laura, who led the processing of both the paper and the Born Digital materials. And then we had three members from the technical, digital side of the library. Ben Ranker was our senior software engineer, we had another software engineer, Pete Hornby, who is also our Apple specialist, and then me. This group included a range of expertise, including traditional archiving processing, research support, preservation, digit research methodologies, a computer programming, content modelling -- the list goes on and on. With such diversity, skill set, and professional perspective, it was vital to especially early in the team's work the roles and responsibilities of each one, as well as the most effective modes of communication for a diverse group like that so we can work effectively together. Deciding to structure the team with co-leads from each division was a remain important and effective decision we made early on, and we also quickly learned that we needed to meet weekly, just to keep the ball rolling. We had a very short deadline, for us, of less than a year to get it out to the public and launched via the MARBL reading room. So we needed to be efficient. And we're going to talk a little bit more a bit later how these weekly meetings impacted the work of the group. And another important step in the team formation was developing and agreeing on a unified mission and a clear set of desired outcomes. Because of the different perspectives and training, backgrounds, professional training and culture, the group first came together with very different notions of what such a hybrid archive might look like when released into the MARBL reading room, through conversations, debates, demonstrations, arm-twisting, et cetera, the group agreed that respecting the hybrid nature of the collection was crucial and working to find an effective balance between donor expectations and researcher needs would be high priorities for our program. Furthermore, the team made a commitment to exploring innovative approaches to giving researchers as much authenticity and context as possible when interacting with Rushdie's Born Digital materials. These missions, respecting hybridity, balancing donor and researcher needs, and providing authentic researcher experiences would drive the many decisions to follow as we begin processing and providing access to the literary personal papers of Rushdie. All right. Back to the [Inaudible] --

>> So next we're going to discuss two case studies within this project that demonstrate the way in which members of the team had to work very closely together in order to accomplish the small goals that kept us moving. And this is just one example of the kind of things we found in the computer when we started exploring. So for this to make sense, though, I want to provide some context regarding the restrictions involved in this collection, the existence of which shaped much of the planning and work flow for this project, in addition to creating many of the challenges that we're facing. As might be expected, much of Rushdie's personal, financial, and legal files are restricted. In addition to all correspondence with his family and literary agents and his family photos are restricted as well. The other major restriction involves all of his journals. Anything that he wrote in his journals post-1989, basically when the file was started, are restricted. He's writing -- he is planning on writing autobiography and kind of wants first dibs on that. He also specified that correspondence from a select number of individuals would be opened only if phone numbers, fax numbers, and home addresses were redacted. And he also made it clear that the Born Digital material would not be accessible via the web. He was not keen on the idea of his files floating around cyberspace, he wanted the material to be available in the reading room only. So I worked closely with Ben and Pete, the programmers, to make sure that only the appropriate finals made it into the final products that were available to

researchers, the database and the emulated environment that Erika is going to demonstrate. Once I was able to view the files I recorded the appropriate information in an access database that was created from the harvested meta data, included file names, directly paths, date created and modified, original computer, et cetera. I added a column valued verdicts. So applying the criteria, the myriad of restrictions I just noted, I assigned each file one of the following verdicts. So as is files can be released for both the emulation and the database. The redacted files would need to be redacted for access. They wouldn't be available in the emulated environment but they would be available in the database. Then we've got the restricted, not available in each one. And then emulation only. These appeared in the emulated only. They didn't go into the database. A good example of the type of files I'm talking about in here, he would get templates, like, I think from his publisher, they just say if the title was format and inside it was blank. So there's no text to search. So those kinds of things.

>> Just to quickly, this category actually marks -- you going to say that?

>> No, go ahead.

>> -- it marks one of those collaborative intersections. Because there were some in our group who are like why are we even keeping these files? Because --

>> Yeah, I wanted to --

>> I'm not naming names. But others of the group felt like -- folks interested in, say, researching how technology and shifts in technology impacted literary production might be interested in this sort of contextual formatting associated with application-driven data and context. The fact that that category exists sort of marks a whole series of collaborative conversations and decision-making that happened.

>> No, that's good. So this is kind of small, but I also -- [Multiple voices speaking]

>> Yeah, it looks small here. I also assigned series and sub series so each file. These were added primarily to enhance the user experience in the database. I felt it was very important, given my background, to provide a holistic and seamless research experience, as the user, who I envision moving from, you know, the paper material to the Born Digital collection, back to the paper. I wanted the ability to organize the files in a similar way in which we organize the paper material. So [Inaudible] series that the researcher would recognize writings and correspondence, and then sub series, fiction, non fiction. So once I had reviewed these files I then delivered the value-added meta data which [Inaudible] my verdicts and series and sub series to Ben. So he let me know what fields and what sort of format was acceptable to him, which informed each future hand-off. For example, I was using [Inaudible] which are alpha numeric strings that uniquely identify a file as my unique identifier or primary key. And that wasn't useful for him. The check stems are too long and complex to serve as permanent identifiers and did not work with the Fedora repository that Ben was working with, so instead we incorporated the Fedora generated [Inaudible] which I think are like, six characters long, as our unique identifiers. Because I was able to add the information to the existing meta data,

users can now not only view the files by looking at his exact file and folder directory, but they can also sort and hone in on the exact type of material they're interested in for their research.

>> So for our second case study we're going to discuss how we've actually provided access to all of this material that Laura so dutifully processed. Creating these points of access for Rushdie's Born Digital materials depended on decisions made at the very point of acquisition, as well as choices made over the course of almost three years. By acquiring the entire computing environment, the whole computers, not simply a data dump, say, of discreet files from selected -- cherry-picked, emery had a rich opportunity to consider a number of access opportunities or approaches. From restricted access to migrated files, say PDF's, for example, to complete emulations of entire operating systems. The approach that we ultimately implemented points back to broader missions of emery's Born Digital archive program, which as I mentioned earlier, respecting that hybrid nature of the materials themselves, balancing donor privacy and researcher needs, and then really attending to sort of authenticity with the researcher experience. So in addition to the very happy for me focus of acquiring the complete computers and the guiding directives of our program's mission, we also made decisions about data capture and preservation that informed the kinds of access points that we later produced. Pete, who I introduced earlier, opted for data duplication over data migration when we first got the machines as a preservation tactic. Which means rather than, say , migrating all the files from these obsolete mid-90's -- Rushdie, he played with every possible word processing program in existence for Mac's in the mid-90's. So instead of migrating all those to more current applications and software, we instead created disc images of each machine, which means we created a bit by bit of each machine, an exact replication. That's the beauty of digital, you can do that. And it offers archives close to complete -- you don't get the hardware, obviously -- preservation of the computing environment. This approach also creates the opportunity for archival staff to safely process and provide access to the material. Data duplication preserves the original while authentically providing exact copies for archival use by staff, and if desired later, researchers as well. So this preparatory work, when combined with our earlier decisions and programatic mission led to us taking a two fold approach to access. We decided to provide users with an emulated environment that provides researchers with a full experience of entering Rushdie's computer, and a search and browser interface that provides full text searching of the user-generated files harvested from the computer itself. So I'm going to try to be very disciplined and take a quick look. Because this material is only available in the reading room, I can't -- there's no web app for it. So we had to create a canned tour of sorts of the material itself. And so if you were in the reading room and you logged onto the researcher work station that we have up there for Rushdie, this is what the interface looks like. We have these little emblems, the first one, what we were calling Rushdie's computer -- right now we only have one computer available. His earliest the [Inaudible] that's the emulated environment, that little computer icon. We next have access to the search and browse interface, which is a collection of derived PDFs from the [Inaudible] and then we have sort of supporting documents, finding aid and help documents. So what did we decide we're starting? These are too long to show in full. So if anyone's interested and wants to see the whole thing, please find us after, we're happy -- they're about ten minutes, there's two of them, amounts to about ten minutes total. But for now I'm just going quickly give you a sense -- this is the emulator booting up. You'll see the emulator booting up in a current operating system.

>> Might look familiar to some people.

>> If you used a Mac in the '90's, this should be like, nostalgic.

>> And that's as big as the screen gets.

>> Because that's the actual size. It's a true emulation. So you know, you log in to the emulation, and this is what you see. You can go to the hard drive, this is Rushdie's desktop. Go into the hard drive, look at his directories and files, he did a lot of work on this computer with ground beneath your feet. So you could say go to a folder, actually see what he has inside a given folder. He loved stickys. Remember stickys? They were awesome. And he created a ton of them with notes for the ground beneath your feet. So this is one of the stickys he created. In addition, I'm going to show you -- quickly show you the searchable data base. [Inaudible] let me go -- 340 -- sorry, this is more emulation. You get it fast forward. My apologies. This is the searchable database. This includes a series -- basically all of the as is user generated files in PDF format, you can do a search, pulls up results with some meta data. There's -- we try to give you sort of a simple view at the beginning, if you call this simple. Not a full meta data view. And we also, if you can see, we're trying to at least include some context by giving you the full directory path so you have some sense of what folder and directory it's in. And then -- we're going to jump ahead just so you can see. This is a -- a more full meta data view. And then it opens up, there's the PDF. So it's in some ways more intuitive, little more familiar because it's more current technology, to be frank. Okay, I must stop. Stop. Yes, [Inaudible] on all that. And that. And that. We'll let you look at that because it's pretty. So -- trying to see where I can skip -- so I think the thing to sort of know about this, as far as collaboration goes, these hand in hand efforts, the passing back and forth of data, working out what the verdict categories are going to be by both the technologist and the archivist resulted in a set of tools that at once provide that full digital context for the files on this earliest machine that [Inaudible] 5400, while also providing access to only the approved content. So authenticity and that balance of donor and researcher needs was realized at once, which we were very pleased with. So emery thought -- if you're wondering, like, why go to all this trouble, which has been asked of us, emery thought it of a special importance to experiment with the researcher access, especially early in this first wave. So we've got to go to findings. Hi.

>> All right. So to conclude -- [laughter] -- we wanted to share some final thoughts about what worked and what didn't work. So what were some of the hiccups? As we went along, we discovered that the division of labor did create some additive work. Instead of me, the archivist, working directly in the front end of the database in creating or revising the meta data in the repository itself, I had to hand that information off Ben who is the midst of building the database at the time. So there was no user-friendly or non programmer-friendly front end for me to enter data. And the more hand offs and the more data duplication is more room for error and increased amount of work. And we also discovered as we reached the end of the first phase that we need to have a further shared understanding of the long term goals. From my perspective, this means we need to ramp up what I call interprofessional education. So you know, me talking to a technologist, you know, well, archivists use series of series to organize [Inaudible] material, and then telling me you can't just write a script for everything. Like, we have to, you know, so, there's that -- but that needs to continue. We're competing with other projects for

digital systems. But I want to end on what worked. Because in this project more worked than didn't, and I wanted to end on a positive. So we had a diverse team of individuals, when the deadlines were set we started meeting on a regular basis, once a week. It sounds like a lot, but I can't tell you how important this is. It really makes you accountable. If you know you have to face your colleague in a couple of days you're going to get the stuff done, you know? And it also creates a sense of comradery. And we all believed in what we were doing. It took a while to get there at the beginning, but we believed in what we were doing, so you have to have that shared goal. It sounds cheesey, but it's just so true, that you have to know what you're working towards. You have to remember the priorities. So there were a lot of times where I was like, oh, can we do this and this. And then Erika would remind me well, if we do that we can't do this. So you're going have to cut something. So -- and so finally, the -- the other aspect of the project that worked really well was that ongoing exchange of ideas and perspectives. You have to go into a collaborative project with a combined sense of confidence in your own perspective, but respect for those of your colleagues. They know things that you don't know, and those things are going to help you get to the end goal. And then finally, speak up when you know something is being overlooked. I think that's really important, especially when archivists are sitting next to technologists, there's that sense of, oh, they know the language and all that stuff. You know things that they don't know, and vice versa. One last paragraph.

>> Quick look into the future, and then we're done. And there's the group.

>> Missing Naomi.

>> Yeah. She was too busy doing so many other things. So now that we know what works we're continuing to advance [Inaudible] at emery. We're current focusing on key components of our infrastructure, and we're also beginning to launch our first phase of user studies, because if we're trying to create an authentic user experience we need to know what the users are actually interested in doing. In addition, we're going to continue processing the rest of Rushdie, we have three more computers and that external drive, and then we have a wealth of other materials and other collections. Much exciting work awaits us at emery and the field at large, and based on our experience with the Rushdie collection we must collaborate or fail. Thanks.

>> That's it.

[Applause]

[Background noise]

>> It's happened before, actually. At the last talk I gave, I had my iPad, I had all my notes on it, and that didn't work very well. So my name is Michael Wilson, I'm a digital collections project manager, at Stanford University Libraries, and we're going to change gears here a little bit. That was a really fascinating presentation that the folks at emery did. And thanks, Jennifer, for putting together the panel. So the name of my talk here is Digital Forensics at Stanford University Libraries, why we have a Fred and why you don't need one. And that should become clear as I walk through the presentation. Just to briefly go over a few things that I'm going to talk about today -- can everyone hear me okay? Okay, good. I'm going to talk very briefly, I'm

going to introduce some of the Born Digital collections that we have, what the imputes was for us to actually building out a digital forensics lab. I'm going talk a little bit about preservation, and some of the issues about retaining prominence and why that's important with Born Digital archives. I'm going introduce you to Fred. My colleagues -- that really stuck for some reason, my colleagues are like how's Fred. But Fred stands for forensic recovery of evidence device. So that's the technology that we're using. I'm going to very briefly introduce disc imaging. Erika and Laura gave a good intro to that, but I'll go into a little bit more depth about what it is and how it works. I'm going to be very daring and actually try to disc image something on the fly here. So bear with me, live presentations and demos are always fraught with issues, but I'm going to take the chance. And finally, I'm going to end with some challenges, some next steps that I'm going to propose to the group here. So briefly introducing our collections, we have roughly -- that we know of -- we have about 18,000 pieces of digital media in our collections. That's a best guess. As many of you know, accounting practices and archives have changed over time. So you know, media back in the 1980's, some of our archivists didn't necessarily think it was important, and no fault on them, but to actually count pieces, track things like formats, elements. So that's a best guess at this point, and we're finding more all the time. And the reason we're trying to find it is we think it's a risk per loss. And I'll go into that a little bit more in detail later. And these are just a sampling of some of the collections that we're currently working on. I think the important thing here is not all Born Digital collections are created equal, they're very different in character and nature. So we have the Steven Cabernet ye collection, which is essentially commercial software, 80% games from '73 through '93, 20% applications. We have Robert Crilly's papers, Robert Crilly the poet. So a lot of his correspondence and drafts of his poems. We have Steven J. Gould's papers. Most of his digital media is drafts of books and chapters of books. And I'll actually show you that at the end of the presentation. We have design files by Peter Koch, from [Inaudible] fine art press, and finally the [Inaudible] project collection, which is a hyper text project that was done in Silicone Valley. So I'm going to kind of boil down why these -- this digital media is at risk, and these three main points I think kind of encapsulate what we're dealing with. The first is media obsolescence. How many people have a drive that can read one of these, the three-and-a-half inch? Yeah, not many of us any more. Well, that's media obsolescence for you. It's just essentially the consumer market has passed it by. So it's very difficult to actually find that media any more. There's lots of it in our old collections, but you know, you go to staples and ask for a three-and-a-half inch drive, it's becoming a little bit more difficult. The second is bit rot, and that's actually what happens when the data on the media itself begins to degrade. I'm not going to go into great detail on how that happens, there's a lot of research in the area. Some of it's -- and there's still a lot that we don't know. But a good example of that is if you put your magnetic disc on top of a speaker you'll lose the data because of the strong magnetic field. And finally, the third challenge is software obsolescence. And that's when you no longer have the software to be able to read -- to actually read the data. And looking at some of the stuff that emery was showing us is a good example of that, where you actually need the applications to be able to look at it in an emulated environment. A good example would be actually finding -- or a good analogy would be finding an Egyptian hieroglyph in 1700 and not having the Rosetta Stone to read it. So I'm going to propose kind of a very basic preservation strategy here for Born Digital materials. This is kind of the basics of what you need to do to really save the data. The first is to get it off the media. That's by actually transferring the bits off the media that's at risk and putting it on to more modern carriers. Whether that's a hard drive, a server, CD, DVD, you name it. Just move it off

the media that's at risk. And finally, back it up. We're all guilty of not doing this enough, but it's good to have more companies, make things safer. There's a couple different ways you can actually do this. And the first one here is probably the one that we're most familiar with. And that's actually using your operating system to make a copy of data on to something else. So an example would be hooking up an external hard drive, finding the file that you want, and actually picking it up and dragging it on to that new medium. The second is to actually create a disc image of it, which is a perfect bit for bit copy of the actual data itself. And I'll go into more detail about exactly what that is. So what happens if you do option Number One, the one that we most often do? Well, the first thing that happens is your creation dates for the object that you copy onto the New Media change for today's date. The second is that you lose a lot of the context. You don't necessarily retain the hierarchy of the files that you transfer onto the new medium when you drag them. Often times, there are files that are hidden or missing, or things that you don't see in your typical browser or Windows Explorer that are critical to kind of the providence of that data. And finally, you know, creation date is an example of file meta data that's changed. But also often times the permissions change as well. And often, there's finger prints on the files themselves that talk about the operating system that actually created them. That can change too. That's not to say that option one isn't an option, it's just those are some of the issues that we're dealing with. So option two is disc imaging. And what that is, as I mentioned before, is a bit for bit perfect copy of all the data, every sector, that's on that medium. The great thing about that is the prove existence of it is retained. You're actually saving all of the same creation dates, et cetera, that come with the file. Permissions stay the same, and you're retaining those finger prints about how the file was created. Often times, there's information about the hardware that's tied in with that, which is important for providence as well. And finally, this is -- I just want to make the point that, you know, I use the word forensics or disc imaging, this is what law enforcement does, so if they seize a computer, they're looking for child pornography, this is what they're doing. And the reason they're doing it this way is because they need to be able to show that chain of custody in a court of law, which is different from what we're trying to do as archivists, but the importance of Providence, it's important in our field too. So now very briefly just going to talk about what we've done to solve this problem at Stanford. We recently acquired two Freds, which are forensic recovery of evidence devices. They're made by multiple companies, most of them in the states. The major customer for these is law enforcement, FBI, and CIA, Department of Defense, et cetera, et cetera. That's the major client at this point. We have two of them, we have a work station which sits in our forensics lab and we also have a laptop that we can take out in the field, if we're going to a faculty member's office or we're going see a potential donor, we can actually go out in the field with it. We've purchased two cat [Inaudible] I put it up there just because the name is so cool. What that -- in a very simple explanation what that is, is that's the older computers had a chip on them, on the mother board, that allowed them to read floppy discs, that allowed them to understand that hardware. All this is, is a specialized card that does the same thing on modern machines. And of course we've got lots of different legacy -- legacy drives, three-and-a-half inch, five and a quarter, tape, zip drives, you name it we've got it somewhere. We have a copy stand in a digital camera to actually take digital photographs of the media itself. Labels are important, it's interesting to see how people label things. It also provides an additional way of identification. And we have sort of ripe lockers, and I'll go into what that is in just a second. And finally, we have forensic software to actually do the imaging process. And this is what Fred looks like. It's -- it's got lots of lights. It's really cool. Up at the top, it's a little bit hard to

see, I realize my photo was a little bit fuzzy. But you can see there's actually right blockers where you would connect the devices, a bunch of local storage drives, and then you might just be able to see down kind of on the bottom of the machine is where we have our -- some of our legacy drives. Like, there's a five and a quarter inch and a zip drive and stuff attached to that machine. Now here the point of the talk is you don't have to have a Fred to do some of the stuff that I'm talking about here. There are uses for having this powerful hardware. But it's@you can do a lot of what we're doing in a much cheaper way. What you need is you need a computer that's able to read the media and run the imaging software. A lot of the software is fully available on the web, there's open source applications and there's commercial applications that are free. And finally, the right blockers are inexpensive. So if you're going to get a computer, what you want to look for is something that ideally still has those floppy disc controllers on the mother board. Most machines in the early 2000's have those. It's not that you have to have them, but it makes it simpler, because you don't need to go out and get a cat weasel and play with trying to connect things, et cetera, et cetera. And you need to retain old drives. As I mentioned, there's many different packages out there. There's [Inaudible] kit which is open source, one I particularly like is called FTK Imager, put together by Access Data. It's freely available, you can go on the web and find it, download it and play with it. But there's literally hundreds of them out there. And finally, right blockers. I mentioned I was going to explain what a write blocker is. A write blocker prevents your computer from writing to the source disc. So it's an important thing to do. A lot of times modern operating systems actually will leave evidence trails on the media that you're trying to capture if you don't use a write blocker. So this is an example of one here for a memory card. I've got a couple samples up here, after the panel is done you can come up and have a look, they're really quite simple. But they prevent were your computer from writing anything on the medium. Floppy discs, if you remember, a lot of them had that little write protect tab, same thing, same function. And they're typically anywhere from $100-300. They're not terribly expensive. So what does this mean? This means that there's really no reason that you can't get started in your archive. The expense is not -- is not great, as I mentioned. You don't have to purchase a Fred. And the write blockers are fairly inexpensive. And as I'm going to show you in a second, the software is fairly easy to use. And more importantly, it's -- it's really not -- you don't have to be a programer to do this. And now I'm going to try the tricky part here of the live demo. So right now I'm launching the imaging program here. So this is FTK imager. I just inserted a memory stick into the computer, and the write blocker -- it's actually a digital camera memory card. So this is what the interface looks like. So it's not so scary at this point, is it? So I'm going to create a disc image. It wants to know what kind. I'm going to copy the whole memory stick. So that's a physical drive. You can see that the software identifies, it's found my hard drive, and it also found the USB reader. It's asked me where I want to save it and what I want to save. I want to save it as a raw image. I won't go into the details of what that is. It means there's no compression, essentially. You get to name it. Add any sort of meta data you want, unique identifier, put in the examiner. [Inaudible] law enforcement, so that's what they have, examiner. Has to know where I want to save it, what I want to call it. And then I press start. And there we go, it's calculating the time it's going to actually take to image this little tiny stick. It's about three minutes for this, 2 gigs. So if you're doing a hard drive it takes a bit more time. But you can go have lunch or do some other processing if you need to. And since I'm impatient, I think what I'll do is show you what it actually produces. I've done this before, so -- so you can see what I'm highlighting here, this is the actual image file itself. Windows doesn't know what it is natively, that's why it doesn't have

any sort of extension on it. But you can see it's 1.46 gigs. So that's the actual image file of everything that was on that little memory card. I've done this more than once, so there's another test that I have. What's interesting here, here's a little output from the software that actually provides meta data of the capture process. So here's the meta data, you saw that screen I was putting in data. Here's a little bit of information about that. And here's the actual technical meta data of the memory stick, if you're interested in that. Most of us aren't. What's really important is Laura and Erika were talking about hash values. What this is, this is a unique identifier that if you move this image to a different piece of media at some point, let's say you want to save it up in a backed up server, et cetera, if you have this hash value you can actually look and if you recalculate the hash it should be an exact duplicate. So it's like a finger print saying this is the exact same thing that we had before. And then there's some acquisition dates, talking about when the process actually happened. And probably the neatest output here and the thing that's quite useful, probably, if you're going to all the trouble to actually capture something, is actually finding out what's on it. And this is an Excel file that's created by the application itself. And as soon as it loads here you'll see it has a list of what's actually there. So we have file name, you'll see a bunch of stuff here, this is because we're actually capturing the partition on the actual memory stick itself. So it's actually capturing the root directory. This is the first partition, it tells you the size, it tells you the creation dates. You'll see some of these dates are a little bit interesting, but actually if you know what's going on, it's not as confusing as it might seem. The actual formatting or partition they have on the memory stick is in fact 16, and that was created in 1980, I believe. But if you scroll down you should see some dates here from the actual images that are on the memory card. You can see -- some 2008. And finally, you'll see it tells you the last date that it was accessed, if that data is retained. An important thing to keep in mind is some of the modern operating systems track that information. If it's a three-and-a-half inch floppy, all they had was date create, they didn't have data access, information is not retained. And if the file was deleted or not.

[Background noise]

>> The application is a little bit of a hog. So here's the final. The final bit here. And I'm kind of throwing this out as a bit of a challenge to everybody in the group. From what I've shown you here, there's no reason you can't get started now. At least move the data off the floppies, whatever it is you have. Create a forensic image, disc image, and your floppies if they die, you're not done for. So that means stop putting them in boxes and actually image them, deal with them as they come in. We don't really have good data at this point on how these things degrade, we might get lucky, we might not. But let's not take the chance. And finally, what I showed here is a very simple way to get started. There are some complications in the process. I think that became clear during Laura and Erika's presentation. There's only an increasing number of media types out there, as the consumer life cycle shrinks the companies release more and more different types of media. I'm sure we knew about the battle between Blu-ray and what was the other DVD format, I forgot. That's a good example. And I think there's -- I think there's a way -- I think essentially moving forward that there's a good opportunity here for certain archives to develop regional centers of excellence. For example, some -- we have computer tapes in our collections, we typically will work with the computer history museum to have them work with our computer tapes. There's ways -- lots of ways to collaborate there to share different types of formats, where for example we might do discs and emery might do Macintosh

computers, et cetera, et cetera. I do think there might be some opportunities for sort of cost recovery on a very basic level there as well. And the final thing here, I want to show you something that's really cool. This is an actual -- this is one of Steven J. Gould's epilogues. And unfortunately, it's a little hard to see. But this is an emulated Word Perfect 2.0 environment that@this is how it looked for him. And that's important, and really, really cool to see. And I'll just read a little bit here. Circumstances demand that this essay receive -- receive the best, most in tell eligible of all epilogues. Though it must occur at my experience. The factual correction of error may be the most sublime event in intellectual life. Anyway, it goes on. But I thought our digital archivist did a good job of it, picking this one thing to show. So thank you very much, and as I said, afterwards you can feel free to come up and ask questions.

[Applause]

[Inaudible comments]

>> Well, I'm going to strip it down even more, I think. I was actually reluctant to take part in this panel today because I -- you know, I saw emery and Stanford and thought they're doing really impressive and advanced work, exemplary work, really. And I don't feel quite up to snuff with them. But I was encouraged to share with you our experience as a small institution with limited resources trying to tackle the same issue, which I hope is where some of you are probably at. So I think this might be one of the least technical presentations you'll see on this topic. I hope what it lacks in technical specifics it makes up for with considerations that any manuscript repository could consider when they're also trying to get started. And getting started is really the hardest part, but as Michael indicated, I think it's absolutely essential to start somewhere, even if that place is not ideal, or even really, I'll say archival. I think too many institutions and archivists, they let the enormity of the task dissuade them from starting everywhere, from doing anything. And this is borne out by, I think, a lot of the studies we've seen on this topic so far. There was a 2008 study published in American Archivist that surveyed 125 archives from across the United States, and like many of you, I expect here today the majority of respondents were from collecting repositories at public and private academic institutions. And of the 125 institutions surveyed, about two-thirds were collecting foreign digital material, accepting them, or planning to. And of those, 80-plus institutions actively collecting them, only 30 had a policy governing those acquisitions, and only 15 had a digital preservation policy. There had been other studies that basically come to the same conclusions. Which to me, that conclusion is I think it's -- first of all, I think it's pretty representative of our profession as a whole, and what it suggests to me is that institutions are accepting this material without any idea what to do with them or where to go next. And this probably means, as Michael talked about, there are a lot of discs found in boxes, in stacks, unmanaged, untouched, bit rotting away, slowly over years. Here's one of ours. [Laughter] there's a lot of discs in that box too. So this is exactly the situation at my institution, the American Heritage Center at the University of Wyoming. It's one of the largest non governmental archives in the country, with over 70,000 cubic feet of archival material. And the focus of our collections is largely the history of the American west, we actively collect in a number of subject areas where we expect to receive a significant amount of Born Digital material. I'll say though, that literally manuscripts isn't one of them. It's not really a strength of our collections. I don't know that we'll be emulating any desktop environments any time soon, or even necessarily accepting computers

as a form of transfer. But we do have a fair amount of Born Digital manuscripts already, and we do have a lot of discs. And we're similar to a lot of special collections departments and academic libraries in that our organization serves as the university archive as well as a rare books library and manuscript repository. But unlike many special collections libraries, the American Heritage Center is not housed within the organizational hierarchy of the university's academic library, and what this means for us is that we don't have at our disposal many technical resources. The technical support we do receive comes from a partnership with the university academic library system department. They are actually very limited in resources themselves. So it's -- we don't get a whole lot of technology out of them. And then we also have our university IT department, but their resources are costly. In fact, any kind of data storage we would use them for would cost us in the neighborhood of $2,000 per terabyte per year. And then there's the issue of human resources. The American Heritage Center is fortunate enough to have a large staff of archivists but only one is current devoted to electronic records issues. And I am also -- my responsibilities also include, as Jen mentioned, managing our digitization program and the center's web sites. I feel that probably many archivists in special collections do not have the staff size we have, and even fewer can probably devote someone's time to a lot of this with so much else to do. Both surveys emphasized that lack of technical expertise and dedicated technical support remain huge impediments to moving forward with Born Digital material. But we as a profession have to start somewhere. If we wait for the resources to move forward, I'm sure we'll wait forever. Yesterday I attended the R B M S cutting to the core session, where the discussion seemed to focus on what can we continue doing and what can we not continue doing. But then there's the issue of Born Digital, where so few have done anything at all, so few have devoted resources or money to it. So I don't even know where that would begin to enter into that conversation, which I think is a serious concern. And if you read some of the seminal articles and reports on electronic reports and Born Digital material you'll find more proposed solutions, I think, than you know what to do with. At the American Heritage Center we have attempted to narrow our scope significantly and make the issue more manageable by focusing first and foremost on this disc in a box problem. So our guiding principle is that at the very least, if you do nothing else, you must separate those electronic files from the physical media. From the physical media it arrived on. Such as the CD's and floppies that Michael mentioned. And we cannot manage or preserve Born Digital materials unless we get the files off the disc, let alone take some very basic archival steps, such as arrangement, description, appraisal, unless we get them off the discs. And you certainly cannot provide access to them. So that became the starting point for everything we started looking at. So in support of this central goal we identified these following sort of general steps. Which I think many organizations, even those with limited resources, to consider implementing. Inventory, existing material, you know, we heard Jennifer talk about the undocumented nature of Born Digital that's coming out in the OCLC report, trying to estimate the number of bytes, because you can't implement a storage solution unless you know how much space you need. Obviously, a transfer of the files from the physical media. But also developing some policies and best practices for future acquisitions. So you know, inventory isn't too crazy, you've all seen one. You cannot I did not know to plan for the storage of Born Digital material unless you know how much you have. So this is an essential first step. And if you're like most archives, you know, that we talked about, you probably have a lot of these discs in boxes. So it's -- it's a matter of sitting down as Michael mentioned, documenting what's on the label, the collection it's from, the media type. And once you do this, you can more easily estimate the number of

bytes. At the American Heritage Center we've scoured our collections in boxes because it's not all been documented so well in our archival management system, and determined that we have probably about 500 gigabytes of Born Digital material on discs, and 350 of that comes from a very recent acquisition of the congressional papers of former Senator Craig Thomas. And it was pretty -- the maximum number of bytes we expected to be in the collection based on the number of discs and the maximum size those discs can hold, it was pretty easy to estimate what the high end of that would be. So you cannot begin -- at the [Inaudible] we're able to estimate, as I said, those 350 gigabytes for this particular question. And once you have that maximum number of bytes, you want to triple it. Loosely speaking. Ideally, you're not going to want three versions of every -- I'm sorry, ideally, you're going to want three versions of every electronic file in your archive. The first copy, as has been mentioned, is the archival copy, and this should be locked down and never accessed, preferably not even by an archivist. And since obviously having one copy of anything is a good way to invite disaster, you're going to want a second copy, you're going to want redundancy. And finally, you'll want a third copy that you and your researchers can use, sort of like the diagram that they ghost-bustered out earlier. And a third copy is ideal, but potentially negotiable if your storage is limited. At the AHC actually, we're not in a position yet to be producing these access file level copies as a regular sort of step. We're doing it on demand or as need dictates. Once you've determined the number of bytes you current have then you can begin to explore storage options. And if your inventory reveals a terabyte or less of data, then you can feasibly support storage of your Born Digital materials for as little as $200, as Michael suggested, the costs are not too exorbitant, and a 1 terabyte external hard drive, for example, is as little as $100, as I said, you'll want redundancy. I think -- I would say that an external hard drive is not ideal in terms of a storage solution, but if you have limited resources, I think it is a step forward from leaving media on discs. Network storage would obviously be a better option because it can provide a higher level of security and redundancy, and typically managed by technology professionals. So start talking to technical staff you might have within your organization, tell them the amount of space you require and see if they can provision any network storage and at what cost. These are the types of things that we've started to ask around about. If organizational technical support is insufficient or unaffordable, then approach university level IT department, perhaps. But again, you can also use external hard drives, if that's the best option you have. At the AHC, we have 3 terabytes of network storage, and it was previously acquired through a partnership with the library and is mostly dedicated to digital surrogates, but it's a place for us to start. It's backed up to tape manually, not actively, automatically, as we prefer, to another server. But it is a place for us to start. I'll call this Frank. This is -- another thing you can talk to your IT department about is just old hardware lying around. You're going to need the media to play it, like I said. What we found -- you know, IT had no use of this machine, it was a retired machine, it had all the drives, so this was free. And once you get to the point of transferring Born Digital material to a storage environment I think things begin to get a little more complicated. If you read any literature on electronic records then you know that there are a throng of requirements that have been developed and written about over the years. And if you attempt to take into account all of these requirements you might feel discouraged from moving forward. If you go the external hard drive example, for example, then you probably aren't going to meet the standards of a trusted digital repository, which you may have heard of. And if you just copy over files from the physical media and do nothing else, then you're not going to meet any of the published requirements for authenticity, which is arguably the most talked about issue with electronic records. My own feeling is that if

you're an institution lacking resources and just trying to move in the right direction, then your plan should seek to comply with the spirit, not the letter of the law. Now the letter of the law might dictate for example the inter[Inaudible] following the [Inaudible] benchmark requirements for authenticity or the implementation of preservation meta data schemas. In reality, though, it's unlikely we could meet the [Inaudible] requirements for discs that sat in boxes for over a decade and probably lack adequate chain of custody documentation. And we definitely don't have the resources to be manually producing XML encoded schemas for all the files on our discs. So complying with the spirit, in my mind, means at least documenting the archival actions you take with any of this material. Even if it's just in a spreadsheet or text file, and includes documenting what we heard about earlier, file authenticity. So I won't explain a check-sum, but as Michael explained, it's a signature. And if it changes, anything in the file changes, even the case of a letter, that signature changes as well and you know it's been altered. So at the AHC we have elected to use an open source tool created by [Inaudible] at the Duke University archives to accomplish both the transfer to storage and the capture of this authenticity information through check sums. The tool is called the data accessor, and it's a very simple, straight forward interface. I know tech people are always saying that, but this is all it is. And I think most archivists would feel comfortable using such a tool. We're actually trying to get some of our archivists, just processing archivists, extension people, using this as well. So far so good. With just a couple of clicks the tool will copy over the entire disc contents to our storage environment without touching the files, while automating the creation of a very basic XML document that lists file names and formats of the disc contents, documents, the folder structure of the disc, and even captures that essential authenticity information. This information combined with our initial inventories will be part of the paper trail we end up generating for all the discs we have in the collection. As I said, you just want to document just sort of what you're doing as you go. Any options you take, any basic steps you take ultimately to include -- though we haven't gotten to this point -- any digital preservation actions you take as well. Accounting for the Born Digital materials already in your collections I think is an excellent first step not only because those materials are likely in dire need of your attention, but also because implementing a process, however remedial, can inform your future collecting activities. The papers of Senator Craig Thomas, which is our largest hybrid collection, as I mentioned, came to us entirely on CD's, DVDs, and floppy discs, after he died in office. At the AHC we are deliberating whether we even want to access material in this way at all. In talking to the staff of current US Senator Mike Enzi, we discovered that his Born Digital archive is about a terabyte-and-a-half, and counting, and all of that data is backed up to his office in Wyoming. So there's no way we want to receive a terabyte-and-a-half of data on CD's. So knowing this, knowing the extent of this material in advance, and knowing that it all exists somewhere that's relatively close to us, we could acquire it without accessing any discs at all. Our preferred method of transfer in this case would be using an [Inaudible] and external hard drive, and since Senator Enzi may serve an additional term or two, we prefer to do regular iterative accessions rather than just all at the end. And it's not just the medium of transfer we want to dictate, but also as much as possible the state of the collection when it comes to us. One thing we noticed with Craig Thomas DVDs in general is a lack of organization of files and folders, which is no great surprise. They look like desktop data dumps, for the most part. Where they simply transfer whatever files were on the computer in a rush. And in many cases, file and folder names are not very helpful in determining the contents of the records. As we do with donations of all paper collections, we like to perform some appraisal with born digital material before we access it, if

possible, and provide some direction to creators on preparing the material for transfer to our archive. I'm sorry? Little behind. In both of these situations we believe preacquisition dialogue with the donor can make a huge impact. Our dialogue with Senator Enzi's office to this point has alerted us to the size of his Born Digital material and allowed us to plan accordingly. And we home further consultation with the office will permit us to make recommendations regarding the use of semantic file or folder names that will make the material easier to work with, both for processing archivists and researchers alike. This has let us begin drafting a brief document for records creators that out line steps they can take with their electronic records to prepare them for transfer to an archive. And the best example I've seen of this is Yale's Author's Guidelines For Digital Preservation, which recommends such actions, as you see here on the screen, save all media, name your files consistently, organize your files. It seems astoundingly obvious. But one explanation of the organize your files reads the management of your digital materials can be enhanced if you handle them in groups and organization them in a logical manner. The structure should be consistent with the organization of any paper records you have or records on other media so that all records related to the same activity or subject or same type can be identified as part of one conceptual grouping. As you can imagine, that will not only assist the creator in managing their own records which might be a welcome recommendation to them, but will also make life much easier on archivists and researchers much later. The Craig Thomas congressional papers have also revealed a few file format that's that we're not sure we can work with. So from this chart we can see there's a file like forest me dot 97, which was unusual, but still recognized by a file format validator as a Word Perfect document. There's another one on there, Senator e-mail list, W.B 3, that was not recognized at all. And may not be possible to avoid acquiring documents in unknown or even proprietary formats. But you can develop a list of preferred formats for common documentary types, and share that list with creators or donors. It can even be elucidated in disposition lists that are shared with donors and congressional collections, for example, or something that we do have disposition lists that we use pretty heavily. Such a list of file formats could also one day form the basis for digital preservation and access plan as well. At the very least, you should certainly attempt to learn of any of these formats to the extent that you can before creating -- before accessing rather than later. It would also be advisable to adjust your donor form to accommodate Born Digital materials specifically. In many ways, the management of electronic records does sort of follow on the heels, I think, of our methods for dealing with paper materials. But digital preservation is obviously not one of them. So at the AHC we added a sentence to our form, our donor form, so that we have the authorization to take any necessary action that we see fit. And obviously, the -- I've -- as I said before, I stripped this way down, and this would not be an ideal solution, I think, in the long term. So keep iterating, start somewhere, and it's not about getting the perfect process from the start, it's about taking small steps toward it over time. So my goal today was to demonstrate that un sufficient technical resources need not deter you from getting started from taking some initial steps or doing some initial planning. It was tempting to turn down this opportunity, as I said, because we are taking I think very humble attempts to deal with the issue. But in preparing this presentation a colleague reminded me of the 2006 presidential address from SAA president, Richard Pierce Moses titled Janice In Cyberspace, Archives on the Threshold of a Digital era, in which he states we need the initiative and drive to dive in and begin working with digital materials. We cannot wait until we have everything figured out. I didn't want to start working with electronic records because I knew there was a real chance of failure. I am enormously grateful to my friend who counciled me early on, whatever we do, we

may fail. But if we do nothing, failure is guaranteed. Discs in a box is guaranteed failure. We may need a dose of reality now and again, capturing snap shots from a geographic information system is not as desirable as preserving the entire system, but it's better than nothing. Capturing a PDF of a spreadsheet means the underlying formulas are lost, but it's better than nothing. Let us celebrate the reality of what we can accomplish rather than bemoan the dream we did not fully realize. I don't propose, as I said, that the approach outlined here is ideal or perfect, but in the spirit of Richard Pierce Moses, I'd argue that the perfect is the enemy of progress when it comes to Born Digital material. So I share this report of our activities hoping it demystifies some of -- some of the aspects that I think are rather non technical, some of the things I can you can do to get started in the hopes that others will make an attempt and share it with the rest of us too. Thank you.

[Applause]

[Inaudible audience comments]

>> Thank you, that was really interesting. I'm Samantha downy, I'm a graduate student at Ben, and I work on the history of the material [Inaudible] so this was a great session. My question is for Ben and Michael. When you get data off of electronic media, you're obviously changing the form still in which it's accessed. And that seems like the useful thing to do, otherwise you're facing obsolescence. But I'm wondering if there's a sense -- a pressure to discard the old stuff? And the reason I'm not asking the emery folks is if you spend so much time chasing Salman Rushdie to get his stuff, then you don't want to toss out the computers. But if it's boxes of old floppy discs, and you realize you have another way to read the data, then is there a sense that, you know, just as libraries are beginning to deaccess books after digitizing, is there a sense that the physical object doesn't matter any more, the old original one?

>> I've actually -- I can answer that, because I've actually had this debate with our manuscripts -- head of manuscripts. Stanford is known for being, you know, fairly wealthy institution. However, we're constantly being pushed off campus and we're running out of space. So there is definitely a pressure, and the questions come up, are we -- should we actually retain them once we get a bit for bit copy of it. I've actually argued that we should save it. And part of the reason is -- is you know, working with manuscripts and rare books, you know, it's an artifact. It does -- physically, it does -- when people touch it, it does mean something. Does that mean that there can't be an argument made for getting rid of them? No, that's a possibility too. Ultimately in the scheme of things the amount of space that 1,000 floppy discs will take up versus, you know, any other archival storage medium is pretty minimal. And one of the things, we never know what technology in the future is going to provide. I'm not sure -- bit copying seems like a great thing now, and I'm pretty confident that's going to stick around for a while. But I'm sure that there's lots of new technologies that can come out that would do interesting things we hadn't thought of.

>> Yeah, physical space isn't an issue for us, especially since if we have room for horse saddles and six-shooters, I figure we can make space for floppy discs. But -- so I don't think that's an issue. But then there is the issue, I think, I don't know if you were mentioning the original electronic file as opposed to the physical media, but I think Michael would probably say to save

both. You can -- you can keep that original however inaccessible. But also create a copy and start migrating that access copy. The only concern I have with that is more the virtual space, which I don't know if that's an issue for emery or Stanford, but it is for us. We -- every bit I count daily.

>> Thanks for a great talk. You all spoke about what you did, like a case study. I'm wondering -- I think emery addressed it, that you're keeping some of the data as raw data. What are your plans for future migration? All those Macs from the days when they were using Motorola, and of course those don't work today at all. So what are your plans, you know, assuming that in ten years everything that we're using today in computers don't work at all. What are your plans for migration?

>> By migration, I'm assuming you mean the -- migrating the disc images? Not really sure --

>> On some platform that's going to work?

>> Oh, the emulations you mean? I mean, that's a real question. You know, are emulations the way that we are using them for access a long term solution. We don't also have an answer for. I think by maintaining -- one of the driving impulses behind organizations like Stanford and emery and the British library and the [Inaudible] and sort of putting a lot of hope in the disc image is that by creating and maintaining exact replicas of those bits, issues of obsolescence are less threatening, just because there are going to be bridges. Unless some catastrophic event happens, which is not possible, but we can't really plan for that. There are going to be bridges. While technology is a very shifting force, very dynamic, at least does create short bridges between the technologies. So as long as we don't keep things in boxes for a hundred years and then say, oh wait, what's this box, we should be able to continue a fluid and authentic traffic of access and preservation. And that -- the disc image is what gives us the hope that we can keep doing that. And so in 60 years, will our reading -- will we have a reading room is a good question. And you know, how will our users be accessing the Rushdie Born Digital material, will it be through these emulations we created? I don't have the answer to that. But I feel like right now -- and this is one of the things we didn't get a chance to say because I went on a tangent I wasn't supposed to go on. But you know, I feel like as long as we sort of stick to our principles of authenticity is just as important with the digital media as it is with the paper, the sense of the art, the aura of the archival artifact that you're talking about doesn't go away, does not evaporate, isn't irrelevant in the digital media, this idea of para text, [Inaudible] is really valuable and that it's our job to promote that, to provide access to that, just like it's our job to provide access to the original paper. I think as long as we attend to that principle, attend to that mission consistently, hopefully we won't be faced with some catastrophic migration issue. So that was a bit of a circuitous answer to your question, and I'm sorry but sort of best I can do.

>> Just tell us when you're going to emulate the emulator.

>> We'll give you the heads up, the meta emulation.

>> It's a really good question, if I can just add one thing to that. There -- if you actually look at who's building emulation software right now, there are other communities out there, there are

other interest groups. The big push that I'm aware of is folks that want to play old games. And there's also a community in the law enforcement as well that's doing this for different reasons than what we would want to do it. But we can't lose sight of the fact there's other groups and other communities out there that need these technologies and are going to continue to need these technologies in the future.

>> Hi. In [Inaudible] book sellers. Two quick things. One to the panel as a whole, and actually both to the panel as a whole. The first, there's an interesting gap that I'm curious about. Emery, you mentioned, you know, clearly Mr. Rushdie gave you a very specific set of guidelines as to what you can do with it. Then you get into digital forensics, which is an area I have a lot of background in. And what I'm curious about, you know, I presume Mr. Rushdie deleted all his porn before he donated it to you. What is your obligation not to turn it over to someone who can do good digital forensics and uncover a layer of a major public figure that he may not want to expose, and are you having those discussions up front, that we won't do this level of work. My other question, completely unrelated to that, is speaking on the archives you're building, I didn't hear anyone talk about remote storage. And you mentioned catastrophic loss, and it's fine to make back ups on site, et cetera. But catastrophic loss requires remote storage, and I presume that's part of what's being done, but I didn't hear that being said.

>> So with Rushdie, we did at the beginning -- and I wasn't involved in these talks, but you can talk a little bit more about it too. But there was concern about what about all the files he deleted. And we came upon an agreement that we would not -- we would not go there. And so that's a -- it's a -- it's an issue that we didn't necessarily talk about, but last night we were talking about it, that idea of that relationship that you have with the donor and the moral obligations and that sort of thing. And Michael can probably touch a little bit more on that. And we do -- and then finally, to ask -- we do have the dark archive, we have it stored on a remote off site location. So sorry we didn't mention that, but yeah.

>> [Inaudible] the repository --

>> We didn't talk about that, yeah.

>> -- but for emery, not just for Rushdie, but right now programatically with Born Digital archives, we have taken the professional sort of ethical stand that we are not going to reclaim data that the creator intended to get rid of. And I just think that, you know, I just -- that was sort of -- not saying that's something we think all archives should do, but that was emery's take on that. We want to have productive, healthy relationships with our donors. I think going through and reclaiming data that was meant to be deleted isn't going to result in that. And then the second question is, yeah, we do -- now, true catastrophic loss is not completely addressed right now for us with Rushdie. We do have an off site location, but it's still within 200 miles. So the sense of like, if we were -- my ideal world, that data would be somewhere out of state. And we don't have it stored -- we can't, because of some of the security restrictions, we can't say hook it up to a distributed preservation network. So yeah, we do have it off site. And if something happened to the CDC, which is about three miles from where our library is, you know, we'd be okay. If something happens to Georgia, we're not so good.

>> Yeah, that's another great question. We have had this debate with our archivists, particularly when it comes to if you do bring in a donor's papers and you find child pornography on it, you are obligated to report that. It's a law enforcement, no if's, and's, or buts. And we don't want to destroy a donor relationship. Not that we want to encourage that, but what a great way to kill that relationship. I think probably the pioneer in this field has been Jeremy Lighten at the British Library. They've done an excellent job, and I have to say we're trying at Stanford to follow in their foot steps and what the emery folks are doing, and that's to create a dialogue with the donors. I know part of the idea of having a laptop and being able to go out is to actually -- a lot of the forensic software, the more detailed forensic software that I didn't show will actually allow a donor to look at some of the old files they have, and really involve them in that process of selection. You really need to have that debate over what's informed consent. One thing to look at if you're interested is the paradigm project did a really good -- really good set of questions that you might ask a donor to kind of construct this conversation with the donor. And I know we've definitely leveraged that work as well. As for off site storage, we're fortunate that we actually do make three copies of all of our data. And since we're on a fault line one copy is sitting in a salt mine somewhere in the middle of America. But you know, it's -- it's -- you know, even if there is a fire, chances are saving things off site is better, obviously better, having a back up on site is better than nothing.

>> One more question here? In the middle.

>> Going forward -- for you who are collecting -- excuse me. Going forward, for you who are collecting living author's composers, whoever. Do you integrate or do you separate them all, separate the information out?

>> Are you talking about when it comes to editions?

>> Well, I'm thinking about when you're talking about Mr. Rushdie is he's obviously going to be writing other books and obviously, the gentleman in Sweden when they're fighting about it right now. So do you -- for your aids and your controls, do you keep them together or have that as a separate decade, et cetera.

>> No, we keep it together. We process our editions and we integrate them into the existing collection so we have one finding aid. Sometimes for a couple of years it might have a series 10, unprocessed editions. But eventually our goal is to get all that integrated. And it would be the same with a successive computer. So when you saw that it was just one computer option, we'll have -- we're going to process the ones we have now and ideally, process the one he's working on literally today. Hopefully we will get that and it will be listed. So --

>> I think both with chronological acquisitions and hybrid acquisitions, paper, we're really trying to create as seamless a research experience as possible. That's sort of a driving mission, to create a really seamless and authentic research experience.

[Applause]