MODULE 13

DIGITAL PRESERVATION STORAGE

Erin O'Meara and Kate Stratton

SOCIETY OF American Archivists

Appendix B: Case Studies

This appendix provides edited interviews illustrating three different approaches to digital preservation storage: a large research library using a hosted storage service, a research library using a locally developed storage service, and a medium sized university using a cloud storage service. The case studies provide a short biography of each interviewee, along with answers to questions posed by the authors of the case study to the interviewees.

University of California at San Diego Library and Chronopolis

Interviewee Roles

Tim Marconi is the IT operations manager at the University of California at San Diego (UCSD) Library. He supervises a team of seven staff who support many IT systems in the library, ranging from everyday systems like email, interlibrary loan, and lending systems; as well as more complex and unique resources such as a DAMS (digital asset management system) and a research data curation program. These systems require a great deal of storage and general computing support.

Sibyl Schaefer is the digital preservation analyst/Chronopolis program manager at UCSD. She ensures that Chronopolis operations function as expected across three country-wide nodes: one at UCSD; one at the University of Maryland Institute for Advanced Computing Studies in College Park, Maryland; and one at the National Center for Atmospheric Research in Boulder, Colorado. The position requires communications management, reviewing documentation needed for things like TRAC/TDR auditing, and tackling questions such as versioning. Chronopolis has a partnership with DPN, the Digital Preservation Network, which is where a lot of their development focus is right now.

Tell us about your repository and the types of digital objects it stores. More specifically, can you describe how UCSD Libraries stores its digital assets for long-term preservation?

We have two different systems that we work with. One is the DAMS, that is the UCSD, Hydra repository. It will be Hydra Fedora at some point, but that migration has not happened. The DAMS is preserved in Chronopolis, so they are two different systems. Chronopolis can

also preserve objects that are not necessarily in the DAMS. So, we talk about both of those and they are different entities.

The DAMS repository is Hydra[14] and eventually Fedora.[15] Our backend is homegrown, using our own repository. We will be going to Fedora eventually, hopefully by the end of 2016. It consists of just under 100,000 objects. It includes research data sets, images, documents, video and audio recordings. All are in the same system and a search can go through everything, but you can also go specifically to assets: digitized versions of collections covering topics such as art, film, music, history, and anthropology through the libraries; or our research data collections, which is through our data curation program if you specifically want to scope for research data.

What is the general extent of your digital holdings and what type of growth patterns are you seeing annually?

The size of our holdings on disk is about 25.5 terabytes. We increased by approximately 5 TB over the last year, due to the addition of research data.

When we transitioned from our completely homegrown systems to Hydra, we put a freeze on ingestion for about a year in 2013-2014. We were pretty static in our preservation assets and our file system was not growing that much. When we launched the new DAMS based on Hydra, we saw an increase—a lot more data more quickly. We opened the floodgates and said "ingest away." I expect it to continue to grow, because the data sets we are getting are much larger. We are not getting that many more collections, but the data that we are getting, especially with regards to research data, are big. We are getting big data, and preserving it is a challenge.

With growth patterns, we have been growing just under 5 TB every year. I expect that to increase, maybe 5–7 or 7–10 TB annually, depending on how large the datasets become.

We are also currently in a holding pattern for our special collections and born-digital records and archives. They will be large. We will need to make some good decisions as to what we want to put in the DAMS and preserve. We have whole hard drives. Do we want to make

14 See https://projecthydra.org/, project homepage captured at https://perma.cc/P9ZH-D3F7.
15 See http://fedorarepository.org/ , project homepage captured at https://perma.cc/PH9D-Z2V7.

all of that available? I don't think so. I think it has to be curated down a bit. The originals will be in the DAMS. We expect that anything born-digital, TIFF files, for example, will be there and will be large.

What type of storage do you use to preserve your digital objects?

Primary storage for preservation objects is an EMC Isilon. The full cluster raw is 2.1 PB, which comes out to about 1.6 PB useable. It runs OneFS, which is the Isilon operating system. It is **networked attached storage** (NAS) that is mounted and run here locally on campus. The Isilon is on premise.

Previously the relationship was one where Chronopolis was a service provider. That has changed because the Chronopolis management has come under the library purview, so now it is an in-house operation. Storage used to be managed by the San Diego Supercomputer Center (SDSC), but now it is managed by the library. The Chronopolis Program Manager also used to be an SDSC position, but is now under the library. It has definitely become more of an in-house service that we run for our own needs, but also to offer to outside institutions.

How did you make the decision to use this storage type?

In 2010, we had "white box" storage, storage that we purchased from a reseller. It was a bunch of servers with a lot of disks, a lot of 2 TB drives. It was a very economical way to get a lot of storage back then. We had 13 or 14 file servers that all had 2 TB drives and they were running RAID 10. We had lots of storage but we were having drive failures in those machines. We had a lot of reliability problems and we were scared about data loss. We did not end up losing any data, but we were worried that we were going to lose some. We also started to run into the upper limits of the actual file systems. The biggest file system we could get on them was 37 TB and our staging group was approaching that all the time. We would keep giving them more space and ultimately they were going to expand outside of the 37 TB we could give them. At that point we needed to find a solution. We did not want to add another server, another map point, with the data in two completely different places for them to manage. Part of the decision was wanting to satisfy our customers, making it easier for them to manage the data that they were putting in and working on. Part of it came from a preservation

standpoint. We were really worried about how these hard drives were failing and if we were going to lose data.

Around 2012, we got a refurbished Isilon, because they are very expensive. It was night and day in terms of management. Now we had one cluster to manage. We had very few drive failures, and when they did, it did not matter because we could lose 36 drives and still be okay because of the way load balancing and node balancing works with regards to data parity. So, we were comfortable knowing that three copies of the data were spread across the five different nodes and we could survive a lot of different failure types. Another benefit is that we got to be on one filesystem—it was a 320 TB filesystem at the time—so we just had to worry about that 37 TB cap that we were running into. Ultimately, we transferred everything to our Isilon and it worked really well for us for three years. Then we were running near the hardware end of life. In 2014, we did a large search again to see if there were any hardware and software solutions out there, and there are many that would meet our needs, but we came back to not having to touch the data again. The Isilon offers a pretty seamless transfer; if you buy another cluster you just join the cluster to your existing one and all of the data moves over without us having to touch anything. We did not have to do any rsyncs; all the data just moved over. That definitely made us want to stick with that. We were able to get a five-year support contract with this one, so we will not have to do this exercise again until 2020, fingers crossed.

Overall we evaluated ten storage vendors to see the different products that could work. There are definitely cheaper ways to do this, but the ease of management—the fact that I do not have to dedicate a storage administrator to this—is worth it. It happens to work well as the storage for Chronopolis, as well. It is harder when we are pricing Chronopolis since it is the most expensive storage that Chronopolis uses. Our other two sites kind of act as subcontractors to us, so they have a set budget that they need to work within to get storage. That is not really dictated by us, but is more what they can purchase with that cost and what they can run in-house without it being cumbersome. Those are the two main qualities we are looking for in storage. But it is all spinning disk, we are not using any tape.

Do you see cloud storage as a part of your storage architecture, and in what role (understanding Chronopolis is distributed storage as opposed to cloud)?

From a system administrator point of view, one of the things that comes up pretty often is that people will say: Chronopolis is our backup. That is frustrating because Chronopolis is not our backup; it is our preservation dataset. Backup, in a system administrator context, is to make sure we have a backup of our data. We would like to get our datasets into a cloud provider, but we have not yet done that because our datasets are large, and we would start with something that is just under 30 TB. Most cloud storage providers want a hard drive instead of transferring via the network. We would like to take advantage of this snow ball from Amazon and put a backup of our operational data in Amazon and do nightly backups, so we could recover from the cloud using that. It would be relatively inexpensive. It is not saving us money, but we need backup somehow and that would provide us with it. Using cloud for backup purposes works well for us. From a preservation aspect, with regards to fixity checking and everything else, that is what we have Chronopolis for.

We have a lot more storage here than we do at the other two nodes of Chronopolis. We need to determine a strategy for what gets distributed across Chronopolis and what does not. The reason is each node does not have 2.5 petabytes—it is around 300 TB at the other nodes. There are cost implications to increase the storage. If we do fill up our Isilon, then we do need to look at a solution to back up our data, and cloud services could be a fantastic way to go for that. But I do want to stress that we would not consider that a preservation service.

What is your opinion on storage media variety and number of copies of your data?

I love the old adage of the 3-2-1: three copies of the data on two different formats and at least one in another location. But, we do everything with spinning disk, we do not have a tape infrastructure. The amount of data we are talking about right now—the investment we would have to make in tape—we could easily get cloud storage for a fraction of the cost. We have no interest in going to tape, but if there was another

format, like if DOTS did exist,[16] that would be nice. Technologies like that sound great, if they end up becoming a reality, but most of these new technologies have never been released for production use. There are a lot of predictions with blu-ray and increased density, but without blu-ray robots and the rest of the needed infrastructure, it's hard to know what would serve us well. I would like to put our data on another format at some point, but I do not have another format that I believe in right now.

I strongly feel that three copies is the minimum; I would like to have more. I do count protection mechanisms, when you are considering RAID that is inherent in storage. I do count those as copies. If you say, "my non-RAIDed copy of this is the same as a triplicate erasure coded system like Isilon or NetApp," I do not think those are apples-to-apples comparisons since there is some protection inherent in these filesystems that should be considered. One copy from one Isilon is not data protection, though. You need to have multiple copies.

For Chronopolis infrastructure, everything we do is on spinning disks, but the systems are all different. So we are not using the same hardware. Even though we do not have variations in type of hardware, there is variation in the actual manufacturer. With regards to the number of copies, we have three distributed copies; but it is likely there are actually more copies since we are using the Isilon here and there is redundancy at the other nodes, too. We do not count those copies. But when we have to do some type of restorative action, that redundant copy is the recommended file to use for the restore before pulling from another node. So there is an assumption there is another copy of the file at each node.

How are you managing fixity across digital objects?

In Chronopolis, we run ACE Audit Control Environment. There is an instance of it running at each of the three nodes. It registers each item at the file level when they are ingested into the Chronopolis system. It also tokenizes them, which is a level on top of that, which protects the object in case anyone changes the checksum. The files are audited every 45 days. As the size of Chronopolis grows, we will be keeping an

16 DOTS (Digital Optical Technology System) is a digital storage media for long-term preservation developed by Kodak. It is in the development stage and not yet available to consumers. http://group47.com/what-is-dots/, captured at https://perma.cc/G64W-WHSR.

eye to make sure that is a reasonable time frame. The 45-day cycle is a pretty aggressive schedule.

How are you thinking about long-term budgeting and costs for your storage needs?

We take the number that we paid for the Isilon in 2015 and divide it by five since that is the supported number of years. The library budgets that amount to their server replacement budget for the next five years, so that at the end of those five years, the library will have the total amount of money that it paid for the Isilon this year to spend in 2020 for storage replacement. The question is: are we sure we will pay the same amount we paid this time and will we need the same amount? No, because there will be growth in collection size and changes in technology, like density differences; but the library will have a set sum of money to then work with. If the library needs to supplement and request more funds, then it will, but right now that is the model for replacement. The library projects for five years; it does not want to project too far out because the landscape changes so rapidly. We want to be able to adapt if we need to.

What are the biggest challenges you are facing with regard to digital preservation storage?

It takes a while to get content from the DAMS into Chronopolis and it has not been done on a frequent basis. Hopefully, we can figure out how to version what we have versus doing wholesale data replacements into Chronopolis.

There is a lot of talk in the digital preservation world that has a tone of a lack of hope: "we're never going to be able to do this," and "it's so hard." It is hard to overcome that and to not get overwhelmed by what could happen in five years: what if every dataset is 40 TB, and what are we going to do if that becomes the new normal? No one really knows, but if we get caught up in the *what if*, we will not be thinking about what we need to be doing right now. It is good to plan ahead as much as you can, but I think lately there has been a lot of defeatism there. Some of the challenge is to maintain hope that it is going to happen. There is a lot of born-digital stuff that we want to preserve and we can do it. Technology is evolving and things are going to get better.

Some of the other biggest challenges are things you already know: there are going to be budgeting challenges with how much things will cost. You will get a lot of people telling you that the cloud is free and why are we not just using it. We are also challenged to explain why digital preservation storage is different—explaining what fixity is and why we need it. Other people may ask: well, Amazon says they hold three copies of everything, so why is that not our preservation strategy? Education can be difficult, too.

Another challenge is file format instability and the dependencies we have around research data that we preserve that has been generated on proprietary software or software that is very specific to a field or even a lab (developed by the lab itself). We have not delved into what it means to preserve data that was generated by that software and/or can only be accessed by that type of software. The DAMS was originally for digitized materials and those file formats are so standardized and popular like TIFFs. But some of the research data file formats I am more concerned about, especially the raw data. Of course that brings up the question of whether we want the raw data versus data that has been processed.

University of Minnesota Libraries

Interviewee Role

Carol Kussmann is a digital preservation analyst at the University of Minnesota Libraries. Kussmann's position is within the Digital Preservation and Repository Technologies department, which is part of the Data and Technology division of the Library. Her role can be interpreted as the intermediary between people within the libraries who want to have digital objects preserved and the digital preservation environment that they have. She tries to figure out the needs of both sides and how her department can address them. This is a fairly new position within the libraries; she has been in the role for two years.

Tell us about your repository and the types of digital objects it stores.

The University of Minnesota includes many repositories that accept the digital content for which we are responsible. Our institutional repository is currently using DSpace, another repository is Drupal based, and another uses ContentDM. We also have digital objects that are not

stored in specific repositories. With the variety of backend systems in use it makes it challenging to understand the full range of materials that we need to preserve. This is compounded by the fact that most of those systems accept almost any type of digital object. DSpace, in particular, allows you to put anything into it. Our DSpace repository includes a wide variety of materials including word documents, PDFs, power points, audio/video files in various formats, GIS files, data files, CAD information, and images in various formats including TIFFs, JPEGs, and Jpeg2000.

Other repositories have a more controlled environment. The materials in our ContentDM repository are mostly digitized material, mostly image files. Pretty much everything that comes into ContentDM has been pre-approved to be digitized and preserved. To assist in our endeavors with other repositories, a list of suggested file formats and associated levels of preservation are shared with depositors. An example of this is the policies for the University Digital Conservancy (UDC), our institutional repository.

The levels of preservation help define how much work we will do on the back end to preserve files. At this point, some file formats are easier to preserve than others. As a whole, we are working on trying to figure out in more detail within our department what file formats we are most comfortable with. We are working with the UDC in trying to build out our policies so we have a wider framework that will work to address the formats we may find ourselves needing to preserve. We need to be able to work with what these repositories take in and understand how we are going to preserve them on the back end.

What is the general extent of your digital holdings and what type of growth patterns are you seeing annually?

We currently have about 300 TB of files/data/information and we expect it to grow at about 100 TB or more per year. Materials come from deposits into the various repositories by individuals, departments, units, or others on campus, but they also come from grant projects that the libraries are involved in.

Most of the repositories allow anyone associated with the university to self-deposit materials. The majority of reports, presentations, and documentation are created digitally and the repositories provide ways for sharing these materials. In the past, reports, presentations,

and documentation of university activities were sent to the archives in paper form but these types of materials are created digitally and can be self-deposited by anyone associated with the university at any time throughout their time here. Because of this, materials are constantly being added.

Grant projects are also a major source of digital files. The libraries have many grant projects that involve creating digital materials, whether that is digitizing analog materials or working on born-digital video projects. Successful grant projects lead to exponential growth of the amount of digital content under our care.

What type of storage do you use to preserve your digital objects?

We are using a combination of spinning disk and tape. The spinning disks are on a Sun/Oracle storage frame and we are using LTO-6 tapes.[17]

Is this storage hosted or on premise?

The spinning disks are onsite and we make the tape copies ourselves. The tape copies are full copies rather than just the changes which are stored offsite. To create full copies we follow a two-week cycle of creating tapes and sending them offsite. Currently we are in charge of the storage equipment, and how and when multiple copies are made; however, this responsibility may change over time.

How did you make the decision to use this storage type?

At this point in time, the server and tape combination was the best solution for us. We looked for high-quality hardware and software. Also, because we were purchasing hardware, we wanted to make sure that what we were purchasing was going to be by a vendor or service that had a good reputation. This is why we chose the Sun/Oracle combination. LTO tapes were also a known product with quality and a good reputation. As with all storage equipment, newer versions of tapes replace older versions of tapes; with LTO tapes, the cycle and support for these different versions is known. We know when new versions of tapes will be released and how long the older versions of the tapes will be supported and by which machines. Having that knowledge and

17 Linear Tape-Open (LTO) is a widely implemented and nonproprietary magnetic tape format. The tape is contained in a cartridge, and at the time of writing seven versions of the LTO format had been released.

understanding of what is being supported, for how long, and when, is good to know because if the vendor you are using is going to stop producing things, you may find yourself out of luck if you cannot use your tapes or equipment any longer.

Do you see cloud storage as a part of your storage architecture if you aren't already using it? If you are using it, what role does it play?

We are not using cloud storage now, and part of it has to do with the amount of content we do have. Similar to the issue of creating full backups and how long it takes with the amount of data we have, sending that amount of data to and from the cloud is cost-prohibitive, especially if you do need to access it more than you would when using it as a dark archive that is accessed only in emergency situations. The price-capacity ratio has not hit a sweet spot yet in terms of price and value for us, so it does not fit our needs now. There may be some things in our care that can go into the cloud or the cloud could be used for certain purposes; but for right now, with the way we are structured, it is kind of an all or nothing thing. With the amount of content we have, it just has not worked for us yet.

What is your opinion on storage media variety and number of copies of your data?

Having content on more than one type of media is good; you do not want to have all of your eggs in one basket. As discussed before we use two different types of media: spinning disk and tape. We have a minimum of three full copies at any one time. Actually, with the way that tape backups are created, there is always a copy onsite, a copy in transition, and a copy in storage, increasing the number of copies we have. Fortunately, we have not had to refresh things or rebuild our infrastructure yet. We have had to pull things off, remove them based on storage getting full, and have had to do some backups after moving things to a different server. Having reliable copies of your data is essential to all of these types of activities.

How are you managing fixity across digital objects?

We monitor or check fixity across our digital objects at the file level using MD5 checksums.

We monitor fixity any time files are being moved—writing the files to tape, for example. If we move files from one server to another we also check the fixity of the files before and after the move. So, anytime we do a move, we check to make sure the move was successful. We also try to do an annual check of everything on the disks that has not been moving, which is a continual process. If necessary we can check the fixity of certain collections or repositories at any time to make sure we check the files on the disks at least once a year.

We are running MD5 checksums per file. We have had discussions with people who wonder why we are not using some version of a SHA-x algorithm. In our situation, we are not worried about people going in and compromising the files and trying to spoof an MD5. We feel that MD5 is secure enough for our purposes and it fulfills the purpose of understanding if something has changed over time or after a file transfer and has the least amount of processing overhead.

How are you thinking about long-term budgeting and costs for your storage needs?

We are well aware of the increase of digital materials at the libraries, which will only continue to grow. To assist with the long-term management and support of these materials, we are looking into how we can share this responsibility with others. For example, the Digital Preservation and Repository Technology department currently does not have the responsibility of preserving published materials in digital form. Some of these and other library materials are currently within HathiTrust.

We are also working with Digital Preservation Network (DPN) to better understand their goals and model and to build partnering relationships around digital preservation efforts. In general we are looking at what is going on with these types of larger networks and looking to see if there are ways to share preservation services and effort. However, with the wide variety of materials in our care we need to understand rights and sharing requirements. Not all of our materials can go offsite. We need to balance these needs if we share responsibility with another organization.

We are also constantly monitoring hardware and software needs and must continue to make solid decisions when purchasing future equipment.

What are the biggest challenges you are facing with regard to digital preservation storage?

The biggest challenge related to digital preservation storage is that the information is not going to stop coming; we are going to keep receiving materials in digital form, and most likely faster than ever. How do we deal with that amount of information?

Think about video files. The amount of storage space required for storing HD video is huge! There are lots of grant project where digital video is being created; do we know what will be kept? The working copies as well as the final edited versions? What are we responsible for preserving?

Just this past year our institutional repository started accepting research data. Currently, we have a smaller research data collection, but that could grow to be larger. How do we address concerns around those collections? There are going to be formats that are going to be more proprietary and unique than what we have seen before, and they could be large datasets.

We are challenged to get people to understand that there are consequences to creating digital files that require long-term preservation. These consequences include a constant need for monitoring, which requires time and resources. It is still very common to hear things like "storage is cheap" and "we can just get more." But think about it: our file sizes are getting larger, which means the space required to save a single file has increased. A digital image taken today on our phone takes up a lot more space than an image taken five years ago on a digital camera. Storage costs themselves may be coming down, but file sizes are going up; eventually you may need more storage for a smaller number of files. Trying to change this opinion—that storage is always cheap—is a challenge because you cannot always just go buy more. Cost, as always, will be an issue depending on what we need to do going forward.

Another challenge is simply the fact that the libraries will be asked to preserve more and more information. To address this, we are working on policies that document what we are able to preserve and what we are required to preserve. These policies will help the libraries communicate to all of our partners what we will and won't be able to do.

In general, it is challenging to put all the moving pieces together. There are a lot of moving pieces and a lot of people that need to work

together and systems that need to work together. Making them all talk to each other can be challenging.

Northern Illinois University Libraries

Interviewee Roles

Lynne M. Thomas is the head of special collections at the Northern Illinois University Libraries and was the co-primary investigator on the Digital POWRR Project during the initial project submission phase for the IMLS grant. Her role in the repository is largely providing content. She collects contemporary literary manuscripts that are often electronic and reside in an Islandora repository.

Jaime Schumacher is the director of scholarly communications at Northern Illinois University Libraries and manages Huskie Commons, the institutional repository. Jaime was the first project director for the initial Digital POWRR Project grant through IMLS, and is now co-primary investigator on the project, now funded by the NEH.

Tell us about your repository and the types of digital objects it stores.

The institutional repository, Huskie Commons, is a DSpace instance, and the place where we collect a lot of the institutional output of NIU, such as peer-reviewed scholarly materials, student scholarly output, departmental publications, theses and dissertations (digitized and born-digital), as well as some datasets. The library's other repository, geared toward special collections in digital form, is a Fedora-Islandora-Drupal instance. It contains items like our dime novels that have been digitized on our Nickels and Dimes digital collections site. It also serves as a dark archive for materials that are still in copyright and need to be preserved but cannot yet be made public, such as the literary manuscripts of science fiction and fantasy literature. The bulk of the digital projects were created over the past ten years through various digitization grants and multi-institution grants. This includes materials from the Abraham Lincoln–Lincoln/Net, a site about the golden age in Illinois. It also includes materials from websites related to Southeast Asia, some regional history center holdings, and University Archives material. There are about five of us in the library that are content contributors to the Islandora repository. Currently, we do not have a staff

member managing the technical side of the Islandora repository, since the digital collections curator recently left the library.

All assets within both repositories are stored in DuraCloud, but we may be moving storage to Amazon Glacier due to cost and shrinking budgets. We decided to use DuraCloud because Fedora and DSpace can synchronize with it, so the process of preserving the objects would be seamless. They currently have been force dumping digital objects into DuraCloud (manually loading files into DuraCloud) and are working toward an automatic synchronization.

What is the general extent of your digital holdings and what type of growth patterns are you seeing annually?

In Huskie Commons we currently have just over 2,600 digital objects occupying 9 GB of space. It is a fairly new repository so what we are seeing right now is a lot of batch uploading of items like our dissertations and theses. We have seen pretty quick growth, but we anticipate that leveling out a bit. In the future we will see a lot more one-off deposits, where faculty members and students deposit their scholarly output into Huskie Commons. Part of that is driven by the fact that NIU recently passed an open access policy that requires all faculty members to deposit their peer-reviewed scholarly work into the institutional repository, thereby making it open access. Increasing awareness of and compliance with that policy will drive a portion of the repository's growth in the future. We are not anticipating the doubling and tripling that we have been seeing over the past couple years as we got off the ground.

The last time we surveyed the Fedora/Islandora repository was when we were trying to pull together our digital preservation case study for the Digital POWRR grant. We figured that we had between seven and ten TB of materials that needed to be preserved. That includes fifteen years of legacy digital projects all together. We didn't expect a ton of growth, but my colleague and I are in the process of applying for a CLIR Hidden Collections grant which would allow us to quadruple the number of dime novels we have digitized at a higher resolution. If we get that grant, then our storage needs are going to at least double very quickly. This will be paid for in part out of the grant in the first few years, but eventually we will have to pay the maintenance fees on that.

We also have another donor-driven project of digitizing sheet music that will be undertaken at some point, which will also massively add to this. We are at the stage now where we are just skirting toward having enough infrastructure to do slow mass digitization of public domain materials in our collections; if we had been at a bigger place with bigger visibility, this would have been done already. So, it is going to grow by a lot.

We have both born-digital and digitized materials. The Regional History Center and University Archives, in particular, have a ton of born-digital university records. My science fiction subset is also a growing area. We are working with the major professional organization for science fiction writers (the Science Fiction and Fantasy Writers of America) and we expect to be taking in their digital files at some point, but they have been holding back in order to organize before submitting them. We have and will continue to get materials that are born-digital as well. They are all held in the same repository and managed together. We don't have separate spaces for born-digital and digitized materials.

The materials in Huskie Commons, the institutional repository, are primarily now born-digital with the exception of the theses and dissertations that are being digitized and ingested. The past three years of theses and dissertations are born-digital.

What type of storage do you use to preserve your digital objects?

For storing and back-up we have a NAS here locally that is backed up elsewhere on campus, and beyond that it is whatever we can get up into DuraCloud. From there we feel like it is in really good hands because they have robust preservation checks and balances going on at DuraCloud. Right now we have the DuraCloud plan that is in both Amazon Glacier and S3, going forward it will just be Glacier. In Rare Books, for the Science Fiction stuff in particular, we have an ongoing practice of using external hard drives, too. We have access copies that we keep here onsite that are locked-down external hard drives for those files, and we update them periodically. They are not technically backups, they are access copies, but they are a safety measure. We have had some issues with backups in the past. In recent memory, we have had some issues with the NAS and the campus-level backups. We are trying to manage those issues while still maintaining file integrity.

How did you make the decision to use this storage type?

For the NAS, it was what was available and pretty much the only option at the time. As a result of the POWRR project we were able to test out and be trained on DuraCloud. We had a very favorable view of it. We found it pretty easy to use and very robust. Once we got our administration's attention by saying, "these backups here locally are not enough; we need to get this original, unique material off of campus and out somewhere else where it can be better taken care of," then we got the funding, and it was pretty much a natural decision to go with DuraCloud, driven by the fact that both Fedora and DSpace could be synced with DuraCloud.

One of the aspects of the initial Digital POWRR grant was to examine and hands-on test both front-end management solutions and back-end storage. The decision making basically came down to: What works with what we already have and is relatively cost effective? That was how we ended up with DuraCloud.

What is your opinion on storage media variety and number of copies of your data?

My opinion is that the more that you can afford, the better, but we cannot always afford as many as we would like. The preservation standard, I think, is about six copies. We are just not in a position to be able to manage six copies. I would be happy with two. I would be ecstatic at three, personally.

I think two is sufficient, if one is in a geographically different location—definitely one offsite. We are in tornado country. Having one of your storage locations be not on campus would be good. If the campus gets hit by a tornado, you want one of the copies of your data to be elsewhere.

Our repositories are smaller, compared to some. It is nice to have a local preservation copy so that if there is a failure, you are able to restore it yourself, rather than jumping through the red tape to get your stuff back from wherever it is—in our case, DuraCloud. It just would be faster and less painful. For us, though, it is a no-brainer to have your stuff offsite in someone else's hands, as long as that someone else is trustworthy.

How are you managing fixity across digital objects?

The short answer is really only sorta-kinda. We are functioning at the good-enough level of digital preservation. We are never going to meet the gold standard, because we are not budgeted or staffed in a way that that would be possible. What we do generally is look for solutions that deal with checksums and automate getting and checking checksums on our digital files, which DuraCloud does. That is the main tool that we use.

Beyond that, I'm not sure if we have implemented using Data Accessioner yet. We are working toward doing it routinely, but I'm not certain that we have gotten that far yet. We are still in the place where what happens before the point of ingest is still a little fuzzy. We are still building procedures for the preparatory stuff, to get things ingested into the Islandora/Fedora repository.

We are running checksums through DuraCloud and we are trusting that if the time comes where we need to restore our local copies of the repository, that those checksums will all match up. We have faith in the system.

The nice thing is that usually you can choose what sort of algorithm is going to be used in creating checksums for managing the fixity. If you can get the right people talking to each other and agree on an algorithm, that will help with making it a bit easier, down the road, so that you're not trying to figure out, "Ok, here's a checksum, but what the heck created it?"

We have gone with MD5. Basically, anytime we move from one system to another we just have to make sure that it is compatible and will work with MD5 checksums. One of the minor advantages to being under-resourced is that you can make choices and therefore not be stuck with analysis paralysis. You can say, "Well, this is what we're doing and that cuts out a whole bunch of other options, because this is what we're doing."

If we needed to move from one system to another, we would recreate the checksums in the new instance and do a full comparison. We have not had to do it yet. Knock on wood.

How are you thinking about long-term budgeting and costs for your storage needs?

In the past, building up to the 7–10 TB of material that is now managed in the one repository, those items were being collected and created without too much thought to the costs of the long-term preservation. One of our current management practices is to think through cost implications when a project is proposed, trying to put real numbers to it, instead of saying "Sure, we can handle a few more gigabytes or a few more terabytes." Both the content providers and the systems folks here at NIU are trying to get a handle on what sort of storage needs we are looking at and what those might cost. Looking at numbers and being honest about that is a big step in the right direction. We are really being squeezed in terms of budget.

When we think about long-time budgeting and costs, not only is it important to have hard numbers—we need to be obnoxious about advocating for digital preservation and paying for those hard numbers, to keep pushing and pushing and pushing. I [Lynne] don't ever shut up about this when talking to my boss, who is currently the interim dean of libraries. When we hire our new dean, I will be obnoxious. We will be asking about that as an interview question and we talk with people across campus about this too, because these are campus-wide projects.

The institutional repository takes in materials from across campus, so this is something that needs to be centrally funded through campus. It can't be just the library's money problem. It has to be the entire campus's money problem. A lot of this is just being willing to be that squeaky wheel. I [Lynne] am a tenured faculty member so I have the ability to be slightly more forward than my colleagues who are not yet tenured or not on the tenure track. I leverage that as hard as I possibly can, because it is important. I am one of those people that will just talk about this as something that is important and explain why; I try to convince people that it is in their own interest to protect the cultural heritage value of library materials. I frame it that way, rather than in terms of risk management. I talk much more about the data losses that will come if we do not solve the problem rather than say, "the library needs money."

We live in boom and bust cycles. What tends to happen is that in the boom years we upgrade our tools or we advance our software—that's when we make the big changes from system to system—because

there is a little more room for outlay. We have to build our systems with the assumption that we are going to have a good run of bust years. Any time we talk about storage costs, having those hard numbers is hugely important because we have to know what our baseline is so that we don't go below it.

What are the biggest challenges you are facing with regard to digital preservation storage?

Filling the technically inclined positions that are currently empty. Those are the folks that do the hands-on moving about of stuff. We turn files over to the people in those positions and then I don't have to think specifically about the preservation function since our processes were designed to support it. That is how we are structured. But, if I am turning my files over and they don't go anywhere, that is a problem. I think making sure that we get to the point where we have a process that is relatively seamless to get our materials ingested and then duplicated out into our storage—that is our challenge. We just have not achieved that yet, and that is really the biggest stumbling block right now. We need to automate that and make it easy somehow.

We focus a great deal of attention on helping people who don't understand this, helping them understand why this is important enough that we have to provide resources to make it happen, talking to them in a way that makes it real for them. Many people face this challenge, regardless of what resource level they have.

Advocacy is always hard. It is always more difficult when you are dealing with folks who are not specialized in the thing that you are specialized in. Then you have a double gap. You have to cross the "advocacy is often uncomfortable" gap and you have to cross the "we are not always speaking the same language even if we use similar terms" gap. When we talk to systems folks, for instance, we talk about multiple copies and digital preservation in a general sense and they say, "Well, we have backups. Isn't that enough?" And we have to cross the information gap to say, "No, backups are not enough. That's barely a baseline for what we are talking about." And we have to cross the gap of "this is also an uncomfortable conversation, because advocacy is often uncomfortable."

When you look at Special Collections and talk to people who are not in this field and say, "I have these medieval manuscripts that I

need to preserve and that's going to cost money," they can understand that. If you then say, "I have these contemporary manuscripts that are in digital form that I need money to preserve," they might respond, "Computers are ubiquitous. They're everywhere, so just make copies." They don't really understand the preservation portion of that. And that makes it harder for donors to write checks for it.

It is a lot easier to convince someone who is willing to write your library a check to buy you something pretty, like a medieval manuscript or a fancy book. It is always easier to convince people to drop a chunk of change on a tangible thing that they can see, something they can have their names on, especially if it is the kind of tangible thing that other people would be envious of. It's a lot more difficult to say, "This digital preservation hub is sponsored by [name of donor]." It is a much harder sell. But in the long run that is the sort of thing that we need to consider because we can't afford not to think about any and all ways to raise money for this sort of thing and find sustainable funding. This is not one and done. Most libraries are modeled on one and done. This is how the whole serials crisis happened in the first place. You buy a book once, it's a purchase that sits on the shelf forever. The cost is sunk and you don't have to repeat it. With serials and with digital preservation, there are ongoing costs and they keep going up. We have to find ways to structure our budget to take that into account.