5-2-2014

Preserving Content from Your Institutional Repository

Wendy C Robertson
University of Iowa

Carol Ann Borchert
University of South Florida

DOI: https://doi.org/10.1080/0361526X.2014.881209

Serials Librarian v.66 (2014):278–288. doi: 10.1080/0361526X.2014.881209

The Serials Librarian, 66:278–288, 2014
Published with license by Taylor & Francis
ISSN: 0361-526X print/1541-1095 online
DOI: 10.1080/0361526X.2014.881209

Preserving Content from Your Institutional Repository

WENDY C. ROBERTSON and CAROL ANN BORCHERT
Presenters

Between institutional repositories and hosting journals, many libraries are becoming responsible for scholarly content in new ways. While Portable Document Formats (PDFs) are the most common formats today, the unique, local, serial content could be in a variety of formats. These items might be digitized text, born digital text, audio, video, images, or multimedia. This article discusses formats that will remain accessible through time (PDF/A, txt, xml) so that contents are not locked into proprietary formats. It will also discuss options for backing up items and associated metadata, including simple backups, off-site storage of files, Lots of Copies Keep Stuff Safe (LOCKSS), Private LOCKSS Networks, and Portico. The article offers suggestions for how your library might best preserve local content.

KEYWORDS *institutional repository, digital preservation, disaster plans, digital archiving, preservation plan*

Digital preservation is a huge topic, meriting entire conferences and specializations. This article provides an introduction to digital preservation for institutional repository (IR) managers and serialists. The 2006 Association of Research Libraries (ARL) *SPEC Kit 292: Institutional Repositories* provides a good definition of an IR:

> A permanent, institution-wide repository of diverse, locally produced digital works (e.g., article preprints and postprints, data sets, electronic theses and dissertations, learning objects, and technical reports) that is available for public use and supports metadata harvesting.[1]

IRs are intended to be long-term homes for intellectual output from one's college or university. People assume this means the contents will be preserved simply because they are in the IR. However, there is far more to preservation than merely adding it to the IR. This is particularly true when preservation is not the primary purpose or priority of a repository.

PRESERVATION REPOSITORIES

IRs generally are not preservation repositories and do not meet the Center for Research Libraries' (CRL) ten basic characteristics of digital preservation repositories.[2] Most IRs have not yet tried to meet all the requirements of a fully trustworthy repository, focusing instead on access to content. For example, a book scanned for inclusion in a preservation repository typically is a series of image files, but IRs usually ingest a Portable Document Format (PDF) access copy instead. While this role for IRs is shifting, this article will not consider an IR as a fully trusted repository. We will be looking at things an IR manager should be aware of and discuss with colleagues when moving in this direction.

Several documents are in existence to assist with the assessment of a current or future repository and provide guidance on what needs to be done to become a proper preservation repository. Alex Ball's "Preservation and Curation in Institutional Repositories" provides a good starting point.[3] Tools like Digital Repository Audit Method Based On Risk Assessment (DRAMBORA) and the Directory of Open Access Repositories (OpenDOAR) policies tool provide additional assistance to create and improve polices and assess a repository, as does the Network of Expertise in Long-term Storage of Digital Resources (nestor) checklist.[4–6] The CRL uses Trustworthy Repositories Audit and Certification: Criteria and Checklist (TRAC) in its auditing and certification of digital repositories, a goal for IRs to work toward.[7] These sources emphasize documentation, transparency, adequacy, and measurability; the documentation and transparency sections are of particular interest for IR managers. For example, can someone easily find the IR mission statement or goals? Does the IR have stated criteria for selection/inclusion? Does the institution regularly review local policies and procedures? Is the funding secure?

FORMAT CHANGES

Storage mediums have changed dramatically over the last several decades. Just because files are stored on a server or in the cloud does not prevent them from being outmoded; after all, the files are still on an actual server. Furthermore, the file formats might no longer be readable. Are files saved from an old Tandy accessible? How about data or a program used in a lab in the 1990s? Existing software might not be backward compatible with old formats (e.g., a Microsoft product file created in the 1990s could be unreadable today). Files may require a computer program that is no longer available. When a company goes out of business and no one retains responsibility for the software, everything running on it could become inaccessible in the future. Programs might be accessible using an emulator, but one still needs access to an emulator. Even if these problems do not exist, the document itself could have had access controls set on it (e.g., password restrictions) which make it inaccessible. This is, of course, inadvisable for an open, publicly accessible archive, but someone might have inadvertently uploaded materials with such controls.

DISASTERS AND BACKUPS

Having a disaster plan is important for an IR and one element on the trusted repository check lists. It is always prudent to have a backup somewhere else, preferably far enough away that it would not be subject to a single disaster. Disasters can take many forms, including fire, flood, tornado, hurricane, earthquake, tsunami, and war. Occasionally, a disaster allows some time to prepare, such as the Iowa City flood in 2008, when the University of Iowa Libraries moved servers out of the building to higher ground as part of Iowa's evacuation. Sometimes a disaster gives no warning, such as the University South Florida's flood, when a drainpipe burst during a rainstorm, right over the local history portion of the journal collection. Computer viruses are yet another form of disaster.

As with home files, a local system administrator should regularly back up the institution's repository. This usually means making regular copies of new files and changed files in case of a server crash. However, do not confuse this backup plan with preservation. Having the data is different from having the data in a form that is accessible and unaltered. The Joint Information Systems Committee (JISC) briefing paper, *Digital Preservation: Continued Access to Authentic Digital Assets,* which is a great introduction to digital preservation, gives a clear explanation of the difference between backups and preservation:

> Disaster recovery strategies and backup systems are not sufficient to ensure survival and access to authentic digital resources over time. A backup is a short-term data recovery solution following loss or corruption and is fundamentally different to an electronic preservation archive.[8]

Backups do form one piece of proper preservation, so confirm that system administrators have not neglected the repository.

OUTPUTTING ALL CONTENT

Ensuring content will remain accessible after repository software changes is crucial. All content and metadata should be able to migrate out of a repository, with no loss of data, including administrative metadata. The items must be clearly identifiable to match the metadata. Assuming the repository uses standards, this should not be a problem; at a minimum, basic Dublin Core should be accessible through The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).[9] The file might not include all fields, such as administrative notes that do not make sense to share with others. All metadata should be extractable in Extensible Markup Language (XML) as comma-separated values (CSV) or in some other open format that retains their structure so that another system may easily re-use them. The exit strategy is also important in case the repository ever needs to move elsewhere due to organizational shifts (e.g., if an institution closes or merges with others, or if funding goes away).

The IR manager should test the output of the metadata and files to ensure they are as anticipated and include any metadata corrections or additions. Make sure the files and metadata are structured and named in such a way to be easily usable. Confirm that Unicode characters are correct, that long fields are not truncated and that all files are there. If there is content on a streaming server or links to data sets, consider the preservation of these items as well. You are the expert on your content, so check the items that might be problematic. Iowa receives quarterly backups of repository data from their vendor, Berkeley Electronic Press (bepress). This quarterly file ensures all their content is on local servers as well as located out of the state with their vendor host. After reviewing the files, the staff found a problem, which resulted in minor modification to their metadata. University of South Florida uses a shell script to do a weekly harvest of the metadata through OAI and curls the information into a file, which they then copy onto a server for backup in another city. Finally, in a linked data world, preserving the metadata for local unique content in an IR is especially important.

Another aspect of an exit strategy and preservation is having persistent identifiers. If the IR migrates to a new system, it should be as seamless for the readers as possible. Persistent Uniform Resource Locators (URLs) can point to the new location so that old links continue to work. A Digital Object Identifier (DOI) could be appropriate for original IR content, but they should not be used for pre- and post-prints or publisher versions of articles in the repository, so a handle system or other Uniform Resource Identifier (URI) could be implemented.

ARCHIVING WEBSITES

While this article is about IRs, website archiving is another way to preserve important institutional output. A local archivist might have identified content that is important to preserve, but which is not appropriate for the repository or which the department or individual does not wish to include in the repository for some reason. In these cases, Web archiving to collect and preserve this content could be appropriate. One can add a project's website to a repository using Web archiving procedures. Maureen Pennock's *Web-Archiving* provides more information on this topic.[10]

The University of Iowa has a subscription to Archive-it, a Web archiving service of the Internet Archive, to harvest and preserve digital collections. Iowa uses Archive-it for journals and newsletters published outside of the repository as well as for journal sites before migration into the repository. They also collect general content from around the University, which is usually of a general and administrative nature, but sometimes includes scholarly content.

Repository content, especially unique content, may be included directly in Internet Archive, as can be done with digitized books. The State of Montana has uploaded documents to Internet Archive, treating it somewhat like a repository.[11]

IR-SPECIFIC CHALLENGES

An IR is a little different from other library digital collections for which the library typically owns and uploads all the content. If the library uploads all the files, it receives them all before the upload. However, the repository copy might be the only version the library has, if authors or editors submitted some of the content or the content was added automatically via Simple Web-service Offering Repository Deposit (SWORD). Repository content is a mixture of born digital and digitized. Even if the print exists, departments might not retain the original print after scanning, such as if a department scans grey literature before submitting it.

The scans could be high quality, preservation scans, or simply access copies. Library staff might have digitized the items, or scanning might have occurred elsewhere, resulting in little library control over this process. If the library has preservation scans, preserving the access copy may not be as important. The preservation needs of the repository copy could also be different when the analog item exists. This mix of born digital, scanned with no print, scanned with print, and scanned to preservation standards creates an extra challenge when deciding what exactly needs to be preserved.

Some repositories add a title page as part of the upload to brand the content, give it authority, and identify it with the repository name, full citation and rights information, which is often lacking on an individual item. However, this inserted cover page has now altered the content, so consider which version is preserved—the actual original, or the one with the cover page. Furthermore, the process that adds the cover page may create a new

PDF out of the cover page and the original file (versus inserting a page into the original PDF) and the new PDF might lose features, such as tagging for accessibility. In this case, it is probably quite important to retain the original file before upload.

FILE FORMATS

It is best to use open file formats when possible to ensure files will remain accessible in the future. However, if people self-submit, the repository manager could have little control over what actually goes into the IR. Unless there are strict technical controls on uploads, the repository will receive content that is not in an ideal format. Widely used formats may not be as big a concern because there will be tools for their conversion in the future. The library will need to make a policy decision regarding how rigidly it will regard the file formats; will the IR turn down content in an unusual format? Since IR managers typically want to remove barriers for deposits and preserve them for the long term, we generally allow inclusion of non-ideal content and have to deal with it after deposit. Data sets are not necessarily a special file format, but they pose extra challenges because the library must ensure the metadata include additional information so that these data will be meaningful to others, such as what was collected, what the fields are, and the settings of instruments. Theses and dissertations could also have supplemental files in a variety of formats, including zipped folders, which potentially could cause problems in the future.

Much of the repository's content is probably PDFs, but it might not be PDF/A, which is the best version for long-term archiving. PDF/A is an International Organization for Standardization (ISO) standard "which provides a mechanism for representing electronic documents in a manner that preserves their visual appearance over time, independent of the tools and systems for creating or rending the files."[12] This format does not allow features, which will hinder long-term archiving, such as encryption, or embedded audio and video. While PDF 1.7 is also an ISO standard, it does not require or limit features, making it inferior for archiving. If the institution migrates content to PDF/A, whoever converts the files should record the change. The bulk of Iowa's content is theses and journal articles not created by library staff; journal editors require training and the Graduate College will need to adopt new standards for thesis creation.

Ideally, the institution has a clear and public policy regarding what types of file format migration it will do and what it committed to preserving. Make sure the depositor understands the scope and restrictions of what the library will do; if the files are in a proprietary format, long term, options are limited for the items. It might be best to keep the original and output it as an open version for the future. As always, if the institution does not have a public preservation policy yet, looking at other institutions' policies can be very instructive. As an example, refer to the University of Illinois' "Preservation Support Policy" and "File format recommendations" Web pages.[13,14]

PRESERVATION METADATA

Caplan's *Understanding PREMIS* introduces preservation metadata, defining them as data to support "activities intended to ensure the long-term usability of a digital resource."[15] She also notes that people not directly involved with digital preservation would find it helpful to become familiar with preservation metadata. This is particularly true when evaluating or implementing an IR.

IR managers should record actions taken in repository management that are important to know about for preservation (i.e., altering or deleting digital objects). Even if the library does not have a preservation system, the IR manager might wish to track such actions as adding a cover page or optimizing a large file for faster download. If the content migrates to another format (to be more open or in a newer version), make it clear what was done to the file. These changes can put the "authenticity of the resource in doubt." Caplan states, "[M]etadata can help support authenticity by documenting the *digital provenance* of the resource—its chain of custody and authorized change history."[16] Some technical information can be extracted from files, such as the name of the program creating the item and the creation date, but IR staff should record this information when it is available. Most Preservation Metadata: Implementation Strategies (PREMIS) event types are for recording actions after an item's ingest into a preservation repository, so an IR manager might need to invent local event types to record actions consistently before this step.

METHODS OF PRESERVATION

There are several ways to handle preservation issues such as bit rot and software or hardware changes. One method is refreshing, which involves transferring data between two types of the same storage medium, particularly storage media that deteriorate like CD-ROMs. Migration is a means of transferring data to a new system environment, by converting them from one file format or operating system to another. Emulating does just what it says: emulates an obsolete software platform or old operating system so that users can still retrieve data. Replication involves keeping duplicate copies of files in one or more storage locations. Because data files can corrupt over time, validating data integrity (also called fixity checking) systematically checks data to make sure there has been no bit rot and that the data have not changed or deteriorated. Maintaining metadata to identify file characteristics, not just the normal cataloging metadata we normally think of, preserves information on content and creation of file, its preservation history, and other technical information.

LONG-TERM PRESERVATION OPTIONS

There are three main systems for long-term preservation options, which can also include Open Access journals hosted in an IR: the Global Lots of Copies Keep Stuff Safe (LOCKSS) Network, a Private LOCKSS Network (PLN), and Portico. Global LOCKSS and Portico preserve e-journal and/or e-book content, whereas a PLN can be a means of preserving the other unique content in an IR.

LOCKSS uses a combination of the methods discussed above: copies (replication) are checked against each other (validating data integrity) to make sure they still match and there has been no data degradation. LOCKSS members select e-journal content in individual LOCKSS boxes that are geographically distributed. LOCKSS preserves the format as well as the content, and acts as a light archive, meaning that if the publisher's site goes down, materials will be quickly available through LOCKSS. Once a library has worked with LOCKSS and/or their IR company or platform to allow access to e-journal content for preservation, it should notify the LOCKSS organization that new contents are available for preservation. If a library makes major changes in the metadata or files, it must notify LOCKSS before making the alterations so that the new copy is not rejected from the system since it does not match the "official" copies that are already there. Not all serials are necessarily appropriate for Global LOCKSS; libraries might wish to omit newsletters or material that is quickly outdated or superseded.

Unlike the Global LOCKSS Network, which preserves the more traditional types of publications, a PLN could include a wide variety of material and file types. Libraries can use this for all of their material in the IR, but a PLN requires at least seven nodes or destinations to have enough copies and a wide enough geographic distribution to make this a viable preservation option. Each member of a PLN should be a LOCKSS Alliance member, and the PLN group will need to set up policies and governance to manage membership and any issues that might arise. Things to consider in setting up policies for a PLN include the length of the initial commitment for membership, how much notice a member should give before withdrawing, how the remaining members would remove data from the PLN for a withdrawn institution, and whether the group needs a governing body or steering committee. The PLN members will also need to determine how to handle embargoed materials and decide if the PLN will be a light or dark archive. Some IRs keep an access copy in the repository, and maintain a preservation copy elsewhere. There is a group discussing the possibility of setting up a Digital Commons PLN, and this group has decided to include only content housed within the Digital Commons system. There are a number of PLNs already in place, and a list is available on the LOCKSS page at http://www.lockss.org/community/networks/.

Portico is a dark archive that preserves both e-books and e-journals, and converts the source files to an archival format. Like LOCKSS, not all serials are appropriate to archive in Portico, and Portico can work with some IR companies or platforms directly to preserve traditional journal or e-book content hosted in an IR. In order for content to become publicly available through Portico, there must be a trigger event, meaning the original publisher no longer can or will offer the content, and the material is not available elsewhere.

DEVELOPING A FORMAL PRESERVATION PLAN

We have discussed why preservation is important and the difference between backups and preservation. Having a formal preservation plan is essential in managing data in your repository long term. There are a number of factors to consider in developing a formal preservation plan, including the organizational and financial commitment, stakeholders, local backups versus long-term preservation, storage needs, roles and responsibilities, data ingestion, and the library's policies on deletion and embargoes of materials.

In writing a formal plan, one of many key factors to consider is the organizational and financial commitment to the library's preservation plan. The library must consider whether it has support from the organization, from what level of administration, and at what level of funding. If support extends beyond the library, it will increase the plan's long-term chances of success. In examining who the library's stakeholders are for preserving IR content, the entities funding the preservation plan would need to be included. Other stakeholders would include the content producers (such as the authors and editors), the persons using the data, the owners of the material (e.g., authors, societies, other university departments), and the IR managers.

We discussed the difference earlier between backups and preservation, with backups being a short-term solution for sudden unavailability of data, and preservation being a commitment to long-term access to material. One thing to consider is the library's justification for preserving data versus having a simple backup. Will the library be preserving metadata, content, software, or all of these? The library must determine who is responsible for preserving IR data, and whether the library will join a PLN or find some other option for long-term preservation.

Storage needs are another consideration, particularly in terms of disk space, software, and equipment. The library should plan for the amount of needed disk space, the future software needs for accessing data, and what equipment will be needed for preservation. Referring back to the stakeholders, the library will also need to plan for who will be responsible for managing each of these three aspects of storing—and later accessing—the data. This will also be part of the discussion on the roles and responsibilities for each person involved, in terms of who is implementing the plan, maintaining the data (and how), and providing support for accessing the material and troubleshooting problems.

How data will be ingested is another consideration. The library needs to determine how data will be ingested in the system for preservation or backup, including: how frequently, in what format, who will validate the data to ensure there is no data degradation, and whether this will be accomplished in-house or outsourced to a third party. For some libraries, it could be easier to fund this effort by outsourcing rather than trying to provide and train existing staff. The library will need to examine how well the organization is (or is likely to be) staffed, and whether there is in-house expertise to accomplish the library's data preservation goals. In short, what level of commitment does the organization have to preserve digital information?

CONCLUSION

Libraries have been working hard to build digital collections, many of which are now housed in IRs. Now that we have built them, we need to focus on preserving what we have in them. Although libraries have long been backing up data and preserving hard copy materials, now we must put these two ideas together to preserve the data. Because software platforms and file types change so frequently, there are many considerations and methods for long-term data preservation. Libraries should do an environmental scan to determine what combination of preservation options will work best for their particular institution.

NOTES

1. University of Houston Libraries, Institutional Repository Task Force, for the Association of Research Libraries (ARL), *Institutional Repositories, SPEC Kit 292* (July 2006): 13, http://publications.arl.org/Institutional-Repositories-SPEC-Kit-292/3 (accessed July 5, 2013).

2. Center for Research Libraries, "Ten Principles," http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/core-re (accessed July 5, 2013).

3. Alex Ball, "Preservation and Curation in Institutional Repositories," Version 1.3 (Digital Curation Centre, UKOLN, 2010), http://www.dcc.ac.uk/sites/default/files/documents/reports/irpc-report-v1.3.pdf (accessed July 5, 2013).

4. Digital Repository Audit Method Based On Risk Assessment (DRAMBORA) (Glasgow, 2009), http://www.dcc.ac.uk/resources/repository-audit-and-assessment/drambora (accessed July 5, 2013).

5. Nestor Working Group, *Catalogue of Criteria for Trusted Digital Repositories* (Frankfurt am Main, Dec. 2006), Urn: de:0008-2006060703 (accessed July 5, 2013).

6. OpenDOAR Policies Tool (University of Nottingham, UK, 2007), http://www.opendoar.org/tools/en/policies.php (accessed July 5, 2013).

7. Center for Research Libraries, Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC), Version 1.0 (CRL, Feb 2007), http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/trac (accessed July 5, 2013).

8. Maureen Pennock, for Joint Information Systems Committee (JISC), *Digital Preservation: Continued Access to Authentic Digital Assets* (November 2006), http://www.jisc.ac.uk/publications/briefingpapers/2006/pub_digipreservationbp.aspx (accessed July 5, 2013).

9. The Open Archives Initiative Protocol for Metadata Harvesting, Protocol version 2.0 (June 14, 2002), http://www.openarchives.org/OAI/openarchivesprotocol.html (accessed July 5, 2013).

10. Maureen Pennock, "Web-Archiving," *DPC Technology Watch Report* 12-01 (March 2013), doi: http://dx.doi.org/10.7207/twr13-01 (accessed July 5, 2013).

11. Chris Stockwell, "How Montana State Library Uploaded Batches of Digital Objects to the Internet Archive," *Internet Archive Forum* (December 29, 2010), http://archive.org/post/340223/how-montana-state-library-uploaded-batches-of-digital-objects-to-the-internet-archive (accessed July 5, 2013).

12. Alexandra Oettler, *PDF/A in a Nutshell 2.0: PDF for long-term archiving* (Berlin: Association for Digital Document Standards e. V., 2013): 5, http://www.pdfa.org/wp-content/uploads/2013/04/PDFA_in_a_Nutshell_21.pdf (accessed July 5, 2013).

13. University of Illinois at Urbana-Champaign, "IDEALS Digital Preservation Support Policy" (2013), https://services.ideals.illinois.edu/wiki/bin/view/IDEALS/PreservationSupportPolicy (accessed July 5, 2013).

14. University of Illinois at Urbana-Champaign, "Preparing Items for Deposit into IDEALS. File Format Recommendations" (2013), https://services.ideals.illinois.edu/wiki/bin/view/IDEALS/SubmissionPrep#File_Format_Recommendations (accessed July 5, 2013).

15. Priscilla Caplan, *Understanding PREMIS* (Library of Congress, 2009), 3, http://www.loc.gov/standards/premis/understanding-premis.pdf (accessed July 5, 2013).

16. Ibid.

CONTRIBUTOR NOTES

Wendy C. Robertson is Digital Scholarship Librarian at the University of Iowa Libraries.

Carol Ann Borchert is the Coordinator for Serials at the University of South Florida.