UMD-NAL Cooperative Agreement

National Agricultural Library: Digital Curation Plan

A "One NAL" plan for managing and preserving NAL digital materials

Ricardo Punzalan, PI
Adam Kriesberg, Morgan Daniels, Kathryn Gucer, Postdoctoral Fellows
University of Maryland College of Information Studies

Contents

Executive Summary

This report presents the observations, findings, and recommendations of the Agricultural Data Curation team at the University of Maryland's College of Information Studies on digital curation and preservation at the National Agricultural Library (NAL). Through sustained engagement at the library involving the PI, postdoctoral fellows, and Masters fellows, we developed these recommendations for NAL to build an integrated and sustained digital preservation infrastructure which takes advantage of its position as one of the United States' National Libraries and positions it to lead the USDA and agricultural community in providing next-generation information services.

The motivation for this report began with the needs of the Knowledge Services Division (KSD) around preservation of digital research data but has expanded to provide a more comprehensive picture and vision for digital preservation at NAL. Given the organizational structure of the library, there is a strong incentive to collaborate across divisions and build a shared culture of digital preservation at NAL. This report seeks to outline a vision for building this infrastructure, from both social and technical perspectives. Following an overview and explanation of the current state of digital preservation at the library, the report identifies the responsibilities around digital preservation for each of the four NAL divisions before presenting software, personnel, and workflow recommendations. These recommendations are highlighted in the bullet points below:

- Work to establish a shared culture of digital preservation at NAL that crosses division boundaries and includes digital library materials as well as scientific data
- Develop and make public a digital preservation policy framework which
- Build upon existing expertise at NAL as well as broad use in digital repositories by upgrading and expanding the use of Fedora Commons repository software to manage research data
- Seek to hire digital curation staff with a range of expertise in technical and social aspects of digital curation and preservation to compliment existing strengths at NAL

Introduction

The USDA's National Agricultural Library and the University of Maryland's College of Information Studies (iSchool) signed a Cooperative Agreement in 2014 with the goal of increasing collaboration between the two institutions, enhancing data curation activities at NAL, and conducting research on curation and related activities at the library and in the broader agricultural community. As part of this agreement the iSchool has placed fellows in positions across NAL beginning during the 2014-2015 academic year, at the master and postdoctoral level. Through a combination of work on library projects, participation in library events, observation of library activities, and our own experience and expertise in digital curation, we present this report as a summary of our findings and recommendations for the future of digital curation at the National Agricultural Library. We took the "One NAL" philosophy of NAL Director Paul Wester, which seeks to eliminate silos between library divisions and units, as our overarching perspective during this process, looking for ways to increase collaboration between library employees and streamline infrastructure to deliver more integrated user experiences. Digital curation activities fit well within the "One NAL" philosophy.

This Digital Curation Plan begins with some background about the library and the current state of digital curation at NAL. This is followed by a discussion of the responsibilities of various divisions in the library as they relate to digital curation and what will be required to maintain and provide access to digital and digitized materials over the long term. Next we present a sample system diagram as an example of how digital preservation infrastructure might work at NAL. The next section includes a comparison of digital preservation and repository software followed by a series of recommendations for NAL, focused on technology, staffing, and workflow. Finally, the report concludes with a series of short sections on supplementary issues which, while not central to the current digital curation goals of NAL, came up during our work at the library. For these final issues, we have given an overview and pointed to additional resources for reference if NAL decides to pursue these options.

Institutional Background

The management, preservation, and provision of access to agricultural data have always been part of the mission of NAL. In "An Act to Establish a Department of Agriculture," the 1862 founding legislation for USDA, the department's commissioner was charged with a responsibility to "acquire and preserve in his [*sic*] Department all information concerning agriculture which he can obtain by means of books and correspondence, and by practical and scientific experiments, (accurate records of which experiments shall be kept in his office,) by the collection of statistics, and by any other appropriate means within his power."[1] As research activity within USDA has increased, particularly in the Agricultural Research Service (ARS), NAL's responsibility to preserve that research output has likewise increased. Traditionally, dissemination of research results took the form of published reports and monographs, with data tables possibly included as appendices. However, the affordances of digital technology have changed the research practices of scientists significantly in recent decades. Sometimes referred to as the "fourth

[1] "An Act to Establish a Department of Agriculture" https://www.nal.usda.gov/act-establish-department-agriculture

paradigm,"[2] data-intensive science and increased reuse of research data have had major impacts on scientific work. Data files can be shared digitally and, through standardization and the application of information management principles, can be made available to other researchers for reproducibility**,** comparison, aggregation, and reuse. This trend is the motivating force behind recent pushes in government and universities to get researchers to share their data.

This report focuses on digital preservation for data, as projects involving scientific data are the core of KSD's projects, but it also considers the broader preservation infrastructure and possible preservation needs of other units within the library. In service of the "One NAL" vision, it is important to keep in mind the similarities in managing different types of digital materials. While this report was initiated by the preservation needs of research data assets at the library, we believe that the technical and policy recommendations presented here are applicable to other kinds of digital materials NAL currently manages as well as other materials it may want to acquire and preserve in the future.

Current State of Digital Preservation at NAL

In Fall 2015, the UMD iSchool team conducted an audit of the Knowledge Services Division (KSD)'s projects and infrastructure to better understand their relationship to best practices and certifications for digital preservation. In particular, the DKAN platform, on which Ag Data Commons (ADC) is built, was evaluated for its suitability as a preservation system. Currently, the system does not support storage and archiving of the files it houses in the Drupal filesystem architecture, making it appropriate for providing access to materials but inadequate for their long term preservation.

We evaluated KSD systems using the Trustworthy Repository Audit and Certification/ISO 16363 (TRAC) standards checklist[3] because the repository's early planning materials state the goal of achieving trustworthy status. The TRAC checklist is widely considered to be a "best practices" standard for digital repository development and management. The checklist contains three sections: Organizational Infrastructure, Digital Object Management, and Technologies, Technical Infrastructure and Security. While the shared ADC and KSD infrastructure scored well on Organizational Infrastructure (particularly on factors like the size of the team working on the system, the definition of a user community and the access-oriented perspective of the project), other areas did not score as well. These areas for improvement include documentation and policy creation, formalization of workflows, digital preservation processes and technical infrastructure.

Through observation of ongoing work at NAL, as well as conversations with staff from across the library's four divisions, we developed a picture of the state of digital preservation efforts at the library in 2016. At this point, NAL maintains a repository (sometimes referred to internally as the "Unified Repository" (UR), "the Repository," or "Fedora") in which NAL Digital Collections (NALDC) materials, special collections materials, and ARGICOLA metadata currently reside. Fedora is a robust system for

[2] Hey, Tony, Stewart Tansley, and Kristin Tolle, eds. 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery. 1 edition. Redmond, Washington: Microsoft Research.
[3] Trustworthy Repositories Audit & Certification: Criteria and Checklist. http://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf

storing and preserving digital collections that is used in multiple large institutions, including the Smithsonian, the National Library of Medicine, and others.[4] This infrastructure has the potential to support a wider range of digital materials than the unified repository currently holds at NAL. Nonetheless, the technology has not been utilized in this manner to date at NAL, either for the collections scanned through the Internet Archive (IA) agreement or the data-centric projects of KSD. We think that a Fedora-based infrastructure could be the technical foundation for a sustainable, long-term digital preservation service at NAL. We will elaborate more fully on our vision for this service later in the "Recommendations" section of this report.

Responsibilities

This section reviews the responsibilities of each of the units of NAL. Building on the theme of "One NAL" this section seeks to underscore the degree to which different activities across the library contribute to the same mission. It also highlights a similar unity underpinning efforts in the preservation of digital and digitized objects, projects that are currently imagined as disparate or only loosely connected at the library.

Knowledge Services Division (KSD)

Currently, KSD houses a number of data curation projects and databases. These systems make use of diverse software tools, manage different file formats, and serve disparate user communities. The Division wants to streamline and integrate its preservation infrastructure so that everything is managed in a centralized location and can be monitored for stability over time.

This report was, in part, inspired by a growing need in KSD for digital preservation infrastructure to support its expanding group of projects and related services. Teams within the division have been working independently to provide preservation capacity for the data in their control, but have been largely unsatisfied with the level of backups or preservation that they have been able to establish and guarantee to users. Ag Data Commons is discussed in some detail below as an example, but it should be viewed as only one of the projects in KSD requiring more robust digital preservation. Trustworthy digital preservation consists of both regular backups (the shorter-term storage of data or software) and longer-term preservation activities, including redundant copies, fixity checks, format migration, and regular monitoring.[5] KSD needs additional support to back up and preserve their data assets. These unique materials have been transferred to NAL with a reasonable expectation that some level of preservation control has been applied to them, an expectation that is not yet met.

Ag Data Commons

The Ag Data Commons (ADC) project is a key element of NAL's data curation infrastructure. It is designed to accommodate data from a range of scientific domains and provide access in a user-friendly environment. It is built on the DKAN platform, an open source, Drupal-based, content management system with data-specific capabilities and has similar architecture to many other NAL websites. ADC is a

[4] For a more complete list of registered Fedora installations see http://registry.duraspace.org/registry/fedora
[5] DataONE. 2012. "Lesson 6: Protecting Your Data." https://www.dataone.org/sites/all/documents/L06_DataProtectionBackups_Handout_FINAL.pdf.

key system for the library and for all of USDA, as it contributes to the agency's efforts to fulfill the 2013 Office of Science and Technology Policy (OSTP) memo on public access to federally funded scientific research.[6] While the NAL's DKAN platform provides access to datasets, it does not have a built-in preservation mechanism for submissions. The system is not designed as a digital repository and stores all files in a single location on a server, rather than in a preservation environment.

As the above suggests, additional work is required to implement a repository infrastructure that provides access to datasets in a user-friendly way through the DKAN platform while also allowing the ADC team to provide public assurances of digital preservation practices. ADC's goal is to be a broad repository for data from across the agricultural sciences. To attract the contributions needed to reach this goal, the repository needs to be more transparent about its capabilities and assurances. Researchers need incentives to select ADC rather than another repository. Pre-eminent publication venues respected by this research community, such as *Nature*, maintain lists of suggested repositories[7] based on their capabilities and reputation. ADC should aim to provide robust enough services to warrant inclusion in such a list.

i5k Workspace

The i5k Workspace is both a repository and a collaborative workspace for insect genomics researchers, only some of whom are affiliated with USDA. While the original genetic sequences themselves are deposited in the National Center for Biotechnology Information (NCBI), the annotations and analyses conducted by i5k users constitute valuable data in their own right, and serve as the basis for peer-reviewed publications. The i5k Workspace requires both backups and long-term preservation of its contents. Because it contains an active workspace, the system requires periodic backups and logs of changes to the database, both for maintenance and administrative purposes. Long-term preservation secures the results of analyses which may not emerge until multiple versions of results are created, which may make their way into the literature or inspire future work in the insect genetics community.

LCA Commons

The Life Cycle Assessment (LCA) Commons project is KSD's oldest ongoing initiative and one of its most complex projects. The LCA Commons system provides access to valuable and unique data, tools, and other resources that can be used to model the inputs and outputs of a wide range of crops and agricultural products. The LCA Commons supports access to the unit process files that form the basis of life cycle models created by researchers studying these agricultural products. Throughout its long history, the LCA Commons team has built successful relationships with groups conducting similar work at the Department of Energy and EPA. As with the other KSD projects, however, the LCA Commons has encountered significant challenges, including those related to metadata and data management and the need for additional preservation action to be taken in order to assure its persistence over time.

Long-Term Agroecosystem Research (LTAR) Network

The Long-Term Agroecosystem Research (LTAR) Network team at NAL provides data curation and access services for a research network of eighteen field sites for data collection across the country. Together

[6] https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf
[7] http://www.nature.com/sdata/policies/repositories

the LTAR team at NAL and the researchers in this network have worked on issues related to standardization, metadata, and workflows so that the library can better ingest and provide access to these researchers' meteorological data. Preservation should be part of these conversations as well, especially due to the unique nature of LTAR data. Because the data reflect meteorological conditions in the field at a given time, they can be compared to historical data from some of the same sites, making easy reuse of these data sets a high (but as yet un- or under-addressed) priority. The LTAR team at NAL has also been working to provide wider access to USDA's geospatial materials through Geoserver, an open source tool for providing access to GIS data. These two sample types of data (meteorology from instruments in the field and GIS data) represent only part of the data that LTAR sites are creating. As with other KSD projects, LTAR data should be incorporated into NAL digital preservation systems. The LTAR team should work with the Information Services Division (ISD) and other stakeholders in KSD to create a preservation workflow for these materials, deciding, for example, whether they should be copied to a preservation server periodically or at the point of ingest into the access/production system.

Information Products Division (IPD)

The Information Products Division (IPD) is responsible for NAL's primary web presence through the library's website (which offers access to Agricola, NAL's online catalog, among other services) as well as NALDC. To ensure that NAL web content is included in the library's digital preservation plans, it is essential that IPD be part of the conversation about the next generation of digital preservation infrastructure at NAL. Furthermore, the division has stated their desire to bring materials scanned and currently hosted by the Internet Archive into the library's control, materials that also must be considered under NAL's digital preservation purview. This goal aligns with other strategic initiatives involving digital materials and further underscores the need to coordinate efforts and define workflows across the library.

Data Production Division (DPD)

The Data Production Division (DPD) is a key unit within NAL in the area of digital preservation. This group conducts digitization of NAL resources and provides access to special collections materials online. These processes are currently integrated with the library's Fedora-based Unified Repository and involve coordination with ISD. As part of a broader digital curation program, these workflows should be integrated with other divisions and additional capability for research data should be added to NAL's preservation program.

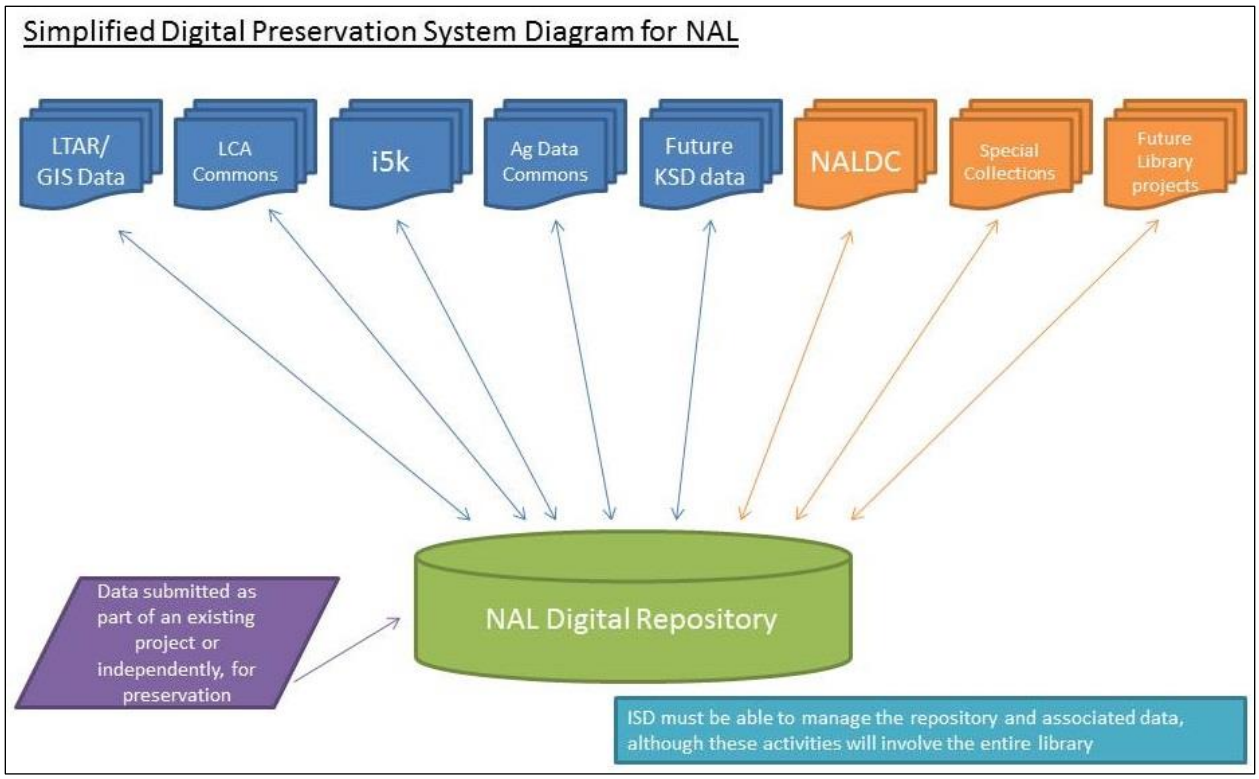
Information Systems Division (ISD)

The Information Systems Division (ISD) is a key organizational component for NAL's future in digital preservation. As the division responsible for maintaining NAL's technical systems and infrastructure, employees of ISD will interact with the other divisions of the library who supply digital objects requiring preservation. The division should be the ultimate home of system administration and related infrastructure maintenance functions for a robust digital preservation repository. As the library moves forward on building a digital preservation infrastructure, ISD needs to participate in the discussion about how to best support the library's preservation needs for a range of materials in digital formats. The development of workflows to ingest, store, monitor, and provide access to digital materials is a key

component of the plan for digital curation at the library and ISD can be a leader in this process, due to the division's insight into the needs and work of the rest of the library. Beyond workflows, ISD also needs to continue sustained engagement with key members of the other library divisions. These individuals should have access to system administration tools as well as reports on preservation system statuses.

System Overview Diagram

The following is a high-level overview of a proposed system architecture for all of NAL's digital preservation infrastructure. The data materials and digital library materials may vary from a technical standpoint but they do need to be ultimately managed by the same team in ISD. As expressed earlier in this report, the idea behind the diagram below is that a single repository will serve as the preservation environment for all NAL digital objects. However, it is important to note that other architectures would still allow the library to fulfill its preservation goals, such as creating multiple repositories.



Digital Preservation Software

The following table is a comparison of fourteen digital preservation software tools, including some already in use at NAL. Various pros and cons are enumerated as well as other variables impacting their suitability for use at NAL. The tools are identified as open source or commercial, other users in the federal government are listed, as are the standards and file formats accommodated by the software according to documentation available online. Finally, we attempted to consider what customization

might be needed to meet NAL's known requirements. While this is not guaranteed to be a fully comprehensive list of preservation and repository software, it does include the major players in the field and provides a set of useful parameters against which these tools can be judged.

Fedora	http://fedora-commons.org	Open Source	DuraSpace (non-profit)	Smithsonian, NLM, others	Yes, used for NALDC	Flexible wrt standards	Many formats possible	Robust, open, scalable	Using web services to pull data from production systems into preservation environment	
DKAN	http://www.nucivic.com/dkan/	Open Source	NuCivic	Multiple; http://www.healthdata.gov/	Yes, Ag Data Commons	Project Open Data	flexible	Integrated with data.gov, Drupal flexibility and design functionality	DKAN currently does not have preservation tools or services built into its software.	DKAN is optimized for data.gov syncing and Project Open Data but is not necessarily ideal for preservation or other curation activities.
Dataverse	http://dataverse.org	Open Source	Harvard University	Do not think so	No	Can accommodate multiple standards including geospatial (unknown re: ISO)	flexible	Establishes link between articles and data in the system	Workflow integration, possible additional metadata work	
DAITSS (Dark Archive In The Sunshine State)	http://daitss.fcla.edu	Open Source	Florida Center for Library Automation (FCLA)	None; used by public university system of Florida	No	Can be configured to accommodate multile standards	Many formats possible	DAITSS does not have a public-facing access system	Would need to be integrated with DKAN or other access system to provide public-facing access to materials inside.	
Eprints	http://www.eprints.org/uk/	Open Source	University of Southampton (UK)	None in US, project based in UK	No	Configurable, can be minimal as evidenced by http://researchdata.bath.ac.uk/policies/metadata.html)	Many possible	Meant to create institutional repositories	Design not so good, we would need to do work in this area.	
Greenstone	http://www.greenstone.org	Open Source	New Zealand Digital Library Project	None in US	No	These are the default metadata standards (others can be added): Dublin Core (qualified and unqualified) - RFC 1807 - NZGLS (New Zealand Government Locator Service) - AGLS (Australian Government Locator Service)	Most standard formats	Can be run as a standalone software tool that can be run offline, via USB stick	Metadata standards supported not adequate for KSD. Additional work would be needed.	Not designed for data.
Knowvation	http://www.ptfs.com/knowvation	Commercial	Progressive Technology Federal Systems, Inc. (PTFS)	Many federal customers, see http://www.ptfs.com/federal-government-customers	No	From website "Supports various metadata schemas including Dublin Core and extended Dublin Core"	Standard	Fully integrated system	Commercial product would be customized by vendor.	Does not seem to be geared for research data
CONTENTdm	http://www.oclc.org/contentdm/en.html	Commercial	OCLC	US Army Corps of Engineers (http://cdm16021.contentdm.oclc.org/cdm/); many universities and state govermnents use this platform	No	Metadata templates available (presumably DublinCore)	Mostly meant for file types that are *not* data, but could be flexible	Full-service solution from OCLC	Would need to make this product better suited for data	Hosted and managed by OCLC
DigtiTool	http://www.exlibrisgroup.com/category/DigiToolOverview	Commercial	ExLibris (ProQuest)	LoC and other fed users are Ex Libris customers. Not clear if they use DigiTool or no	No	Generic "Support for library standards"	Open, but geared towards library content rather than data	Integrates with other Ex Libris products	Vendor would have to scope and complete any customization	
Preservica	http://preservica.com/	Commercial	Preservica	No US federal customers but multiple state archives/libraries, and national archives in EU	No	Flexible standards	From website "over 800 different file formats – including objects embedded in complex container formats such as ZIP, WARC and MS Office"	Commercial product means access to professional suport and development teams.	Vendor would have to scope and complete any customization	In use in government settings in the US. Widely used and liked in state archives/records community. A true contractor digital preservation solution
VITAL	https://www.iii.com/products/vital	Commercial	INNOVATIVE INTERFACES INC.	None highlighted on company site	No	Can use a variety of standards	from website, can store any file format	Built on Fedora architecture	Lots	Not really designed for data, tailored for digital libray
Digital Preservation Software Platform	http://dpsp.sourceforge.net	Open Source	National Archives of Australia	No US, project maintained at National Archives of Australia	No	Flexible	Xena software converts content into open standards. List of formats here http://xena.sourceforge.net/help.php?page=normformats.html	Modular, the platform has multiple components which can be customized.	Does not seem to support the database formats in use at KSD. Customization might be necessary to accommodate i5k formats.	Seems too small and minimalistic to scale with NAL or to integrate into our workflows. This platform is a group of smaller pieces of software into a suite.
Archivematica	https://wiki.archivematica.org/index.php?title=Main_Page	Open Source	Artefactual Systems	None, but used in multiple US university libraries	No	Supports common standards.	Supported formats can be explored here https://wiki.archivematica.org/Media_type_preservation_plans.	Existing evidence of ingesting datasets via web services	Unclear support for database ingest and management workflow. Would need to look into futher if this option is to be pursued.	Open source so future development unclear.

Recommendations

Technical solutions alone will not be sufficient for building a digital preservation environment at NAL. In order to build and maintain such a system over time, the entire library must take on preservation and its associated activities as critical elements in NAL's mission and future success. The following are the primary areas where we suggest that action be taken to strengthen the library's ability to perform digital preservation services now and into the future.

Software

NAL currently has a Fedora 3.x installation maintained as the "Unified Repository" (UR) for the library. As stated above, Fedora is robust software, with the capacity to manage a much broader range of digital formats and materials than NAL currently entrusts to it. The UR team plans to upgrade the existing system to Fedora 4.x in the future but this will be a complex operation as the software has changed significantly in the new version. We suggest that NAL build a Fedora 4 repository and use the opportunity to both learn the new software and integrate research data into the digital preservation infrastructure of the library.

Personnel and Staffing

While the content of this report is intended to provide NAL with observations and recommendations to strengthen digital preservation efforts without specific requirements for personnel, there are nevertheless some skillsets[8] and position descriptions which would fit into each of the branches that currently exist within KSD: Scientific Data Management and Scientific Data Engineering. The descriptions that follow are not meant to directly apply to job descriptions for NAL but rather describe the skills we feel are necessary for the next phase of digital curation at the library.

A member of the Scientific Data Management team should combine information science, data curation, and project management skills to effectively steward research data across its lifecycle. This individual should have an understanding of the technical tools and infrastructure necessary to support digital preservation but need not be a programmer. This person should be familiar with metadata standards for data and information access such as Dublin Core, Project Open Data, Ecological Metadata Language (EML), ISO19115, and PREMIS. They should also be familiar with technologies such as JSON and XML, which are used for metadata management. In addition, a Data Curation professional will be familiar with digital preservation principles such as Open Archival Information System (OAIS) and the Digital Curation Center's Curation Lifecycle Model.[9]

A member of the Scientific Data Engineering team should possess technical skills to support curation and preservation efforts across KSD in addition to an understanding of informatics and the ability to collaborate with ISD. This person should possess technical skills including (but not limited to):

- Programming languages such as Javascript, Ruby on Rails, or Python
- Software such as Fedora Commons, Hydra, Solr, or other digital repository packages
- APIs and RESTful architecture to build web and backend services

[8] The issue of data curation skills and job descriptions has been tackled in the literature for some time. For more, see: Pryor, Graham, and Martin Donnelly. 2009. "Skilling Up to Do Data: Whose Role, Whose Responsibility, Whose Career?" International Journal of Digital Curation 4 (2): 158–70. doi:10.2218/ijdc.v4i2.105; Committee on Future Career Opportunities and Educational Requirements for Digital Curation. 2015. Preparing the Workforce for Digital Curation. Washington, D.C.: National Academies Press. http://www.nap.edu/catalog/18590; Palmer, Carole L, Cheryl A. Thompson, Karen S. Baker, and Megan Senseney. 2014. "Meeting Data Workforce Needs: Indicators Based on Recent Data Curation Placements." In iConference 2014 Proceedings. iSchools. doi:10.9776/14133.

[9] http://www.dcc.ac.uk/resources/curation-lifecycle-model/.

- Database schema
- Technologies used to manage metadata such as JSON, XML, and RDF
- Geospatial technologies such as ArcMap or open source tools
- System administration skills

In addition, a successful technical member of KSD will have experience working collaboratively and interacting with a wide variety of stakeholders, from scientists and system administrators to project managers, library administration and users.

Having a set of dedicated data curation staff members would help KSD coordinate efforts and manage data over time. These people could be primary liaisons between KSD and ISD and could have access to reports and analytics from all KSD projects. They will have a view into the preservation and curation needs for KSD's materials and also be able to anticipate how to expand and incorporate new data into NAL's infrastructure.

Workflow and Policy

One of the primary issues identified in the original TRAC certification analysis of Fall 2015 was the lack of workflow documentation for digital preservation projects. Robust workflow documentation will provide a centralized resource for preservation information within NAL, help staff understand how digital preservation activities in their unit pertain to NAL as a whole, and increase user confidence in the library's digital preservation capabilities. To this end, NAL should implement the following changes related to workflow:

- The library should maintain a unified workflow for all digital objects which reside in the preservation environment, whether they be data, digitized materials, or born digital USDA publications.
- The responsibility for running and analyzing reports from the preservation system should rest with one or two individuals who have a broad view into the range of digital objects in *all* NAL systems and are properly permissioned to directly view files in the repository. This will ensure that the overall environment can be monitored and any potential issues be quickly identified.
- Explore the possibility of streamlining or otherwise linking the submission tool for NALDC and Ag Data Commons. Give users an opportunity through the NALDC interface they already know to deposit data in ADC, or indicate where that data is located so that ADC can generate metadata to point users to it.

In order to integrate the digitization and preservation workflows across the library, a committee should be established to set these tasks and communicate the needs of each NAL division. Each division has a key interest in the evolution of digitization processes and infrastructure; these interests should all be represented in the planning phase and every effort should be made to establish clear processes for digital objects as they move through the pipeline from acquisition to description/metadata to storage to access. While the workflows will be different for analog library/textual materials that are digitized, born digital USDA materials, and data, coming together to design these workflows is key to building a shared culture of digital preservation across NAL.

The TRAC standard outlines the minimum required policy and workflow documents[10] necessary to achieve certification. While obtaining TRAC certification is not an immediate goal of the library, the list of documents is very useful to get a sense of what the most robust digital repositories have in place. Some of these policy documents have been mentioned earlier in this report, and not all of them are directly applicable in a federal context, but they are nevertheless useful. Of particular importance are the policies and procedures (workflows) around ingest, digital preservation, and access. In

[10] http://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf. For complete list see Appendix 3: Minimum Required Documents, p. 81.

addition to these documents, other library-wide policies such as a publicly facing mission statement and publicly accessible collections policy would be helpful as framing for the secondary policies listed below.

- Contingency plans, succession plans, escrow arrangements (as appropriate)
- Definition of designated community(ies) and policy relating to service levels
- Policies relating to legal permissions
- Policies and procedures relating to feedback
- Financial procedures
- Policies/procedures relating to challenges to rights (only if likely to be needed)
- Procedures related to ingest
- Process for testing understandability
- Preservation strategies
- Storage/migration strategies
- Policy for recording access actions
- Policy for access
- Processes for media change
- Change management process
- Critical change test process
- Security update process
- Process to monitor required changes to hardware
- Process to monitor required changes to software
- Disaster plans

Internet Archive Scanning

As part of the agreement between the Federal Library and Information Network (FEDLINK) and the IA, scanning of NAL materials has been ongoing for multiple years. However, currently the library does not host the scanned files on its servers or provide access to them directly using a USDA access system. At the library-wide staff meeting in July 2016, the Information Products Division stated its desire to bring NAL materials from IA into the control of the library. This goal should be folded into the broader conversation around digital preservation at the library. Specific workflows and automated scripts for identifying, downloading, creating metadata as needed, preserving, and providing access to IA materials should be developed, along with methods for library employees to initiate the ingest of IA materials into the NAL preservation environment/repository.

Related Issues

In addition to the development of core digital preservation infrastructure, a number of other issues related to digital curation may become important to NAL in the future. These are briefly identified and outlined below.

Data Curation Training and Outreach

As a result of the 2013 OSTP memo and other efforts of both the Obama administration and the scientific community, there is an increased focus on data management and curation efforts for federally funded scientific research. [11] NAL is well positioned to provide leadership and guidance to ARS and other USDA researchers due to its stature and current responsibilities around ARS research outputs (papers in NALDC/PubAg). As the National Institute of Food and Agriculture

[11] https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

(NIFA) and ARS require data management plans as part of research proposals, and as NAL systems such as Ag Data Commons increase their capacity for agricultural research data, NAL should consider developing Data Management Plan workshops or training activities to offer publicly. These types of services are widely offered at university libraries[12] and could be adopted as a model for NAL, tailored to the needs of USDA and agricultural researchers. This outreach would help spread the word about NAL's preservation and curation services as well as educate researchers about best practices for data curation and research management. The focus of these training workshops does not necessarily have to fall only on the mandates and funder requirements but could speak to broader best practices around data management by researchers and principle investigators (PI). In addition to workshops, library staff (KSD staff/project managers would be best suited for this type of work) could make themselves available to ARS researchers or other USDA employees for consultative meetings about data practices with an eye towards preservation. Using existing best practices combined with an understanding of NAL's capacity to handle datasets, databases, and other tools, these meetings would benefit researchers as well as NAL through increased visibility among the USDA research community. These types of activities would further reinforce the importance of data curation across the research lifecycle and hopefully improve the quality of submissions to repositories like Ag Data Commons.

Additional Resources

- Carlson, Jake, Lisa Johnston, Brian Westra, and Mason Nichols. 2013. "Developing an Approach for Data Management Education: A Report from the Data Information Literacy Project." *International Journal of Digital Curation* 8 (1): 204–17. doi:10.2218/ijdc.v8i1.254.
- Johnston, Lisa, Meghan Lafferty, and Beth Petsan. 2012. "Training Researchers on Data Management: A Scalable, Cross-Disciplinary Approach." *Journal of eScience Librarianship* 1 (2). doi:10.7191/jeslib.2012.1012.

Web Archiving

As part of a broader digital preservation program, NAL should consider developing capacity to do web archiving, which is the process by which content from the Internet is captured and packaged for long-term preservation in a digital archival environment. The most well-known web archive is probably the Internet Archive's Wayback Machine,[13] but many other institutions are beginning to do invest resources in archiving web content. Web content is collected via crawler software and packaged for further use or storage. The standard format for web archiving is WARC, formalized as ISO 28500:2009.[14]

A number of software solutions exist for web archiving activities, both through paid and open source models. The Internet Archive has a hosted web archiving platform called Archive-It.[15] Customers include college and university libraries, as well as federal institutions such as the NLM, DoE, Department of Labor, and the Federal Depository Library Program. Open source software to do web archiving includes Heritrix,[16] also built by the Internet Archive. Another open source tool is Wget, which can download content from the web.[17] Providing access to archived web content is a particularly significant challenge, with few exemplars to consult. For one such template, see the UK's Web Archive at http://www.webarchive.org.uk/ukwa/.

Web archiving projects should be evaluated relative to NAL's mission and collection policy, similar to other projects. Potential areas of collecting priority include all unique digital information products and projects produced by various units of the library and legacy websites given current efforts to migrate into a new web design and architecture.

---

[12] E.g. University of Maryland's offerings http://www.lib.umd.edu/data/dmp
[13] Wayback Machine - https://archive.org/web/
[14] Information and documentation -- WARC file format http://www.iso.org/iso/catalogue_detail.htm?csnumber=44717
[15] Archive-It. https://archive-it.org/
[16] Heretrix. https://webarchive.jira.com/wiki/display/Heritrix/Heritrix
[17] Wget. https://www.gnu.org/software/wget/

Additional Articles

- Antracoli, Alexis, Steven Duckworth, Judith Silva, and Kristen Yarmey. "Capture All the URLs: First Steps in Web Archiving." Pennsylvania Libraries 2, no. 2 (2014): 155. Retrieved from http://palrap.pitt.edu/ojs/index.php/palrap/article/download/67/370.
- Niu, Jinfang. 2012. "An Overview of Web Archiving." D-Lib Magazine 18 (3/4). doi:10.1045/march2012-niu1.

Software Preservation

Software preservation, while related to general preservation for digital objects, has its own set of challenges and technical needs, which are distinct from preservation for more static objects. If this is something NAL is interested in pursuing, there is ongoing work at other federal agencies on this topic, including the Library of Congress's Digital Preservation unit[18] and the National Institute of Standards and Technology's National Software Reference Library.[19]

Additional Resources

- Software Sustainability Institute. https://www.software.ac.uk
- https://blogs.loc.gov/digitalpreservation/2012/11/preserving-exe-a-short-list-of-readings-on-software-preservation/

Conclusion

In this report, we have outlined our observations and recommendations for NAL's digital preservation efforts. These efforts represent a combination of our best understanding of the library's current operations and our knowledge of the wider practices in the digital preservation landscape. While we believe that these actions and recommendations provide a way forward for NAL on digital curation, we recognize the investment of effort, time, coordination, and financial expense that this plan suggests. Given the history of NAL and the USDA has in collecting, preserving, and providing access to agricultural materials, we see digital preservation as an opportunity for the library to expand its role at the forefront of the agriculture information community.

To achieve the goal of an integrated workflow and software environment for digital preservation of library assets, NAL needs to take a unified approach to its collections. Historically, units at NAL have been siloed in their work and not engaged with each other regularly. In order to built a robust and sustainable digital preservation program at NAL, these units need to collaborate more closely and build workflows, systems, services, software, and hardware that work together while accommodating the users and uses of different types of digital content.

Brian Lavoie and Lorcan Dempsey outline a conceptual and social vision for digital preservation work in their 2004 article "Thirteen Ways of Looking at...Digital Preservation."[20] They define preservation not as a set of technical activities but rather as a philosophy that drives work across the library. Below are eight examples of ways to think about digital preservation.

1. Digital preservation as an ongoing activity
2. Digital preservation as a set of agreed outcomes
3. Digital preservation as an understood responsibility
4. Digital preservation as a selection process

[18] For more information see http://www.digitalpreservation.gov/ and http://www.loc.gov/preservation/
[19] National Software Reference Library (NSRL) http://www.nsrl.nist.gov/
[20] Lavoie, Brian, and Lorcan Dempsey. 2004. "Thirteen Ways of Looking at...Digital Preservation." D-Lib Magazine 10 (7/8). doi:10.1045/july2004-lavoie.

5. Digital preservation as an economically sustainable activity
6. Digital preservation as a cooperative effort
7. Digital preservation as a complement to other library services
8. Digital preservation as a public good

Each of these ideas relates to the mission and ongoing work at NAL, in particular the "One NAL" approach to library services, which aims to remove silos between divisions in the library and streamline services offered to users.

In addition to Lavoie and Dempsey's philosophical approach to digital preservation, Phillips et al. provide another window into a set of community-accepted practices associated with preservation activities. Known as the Levels of Digital Preservation, we have reproduced below the chart explaining the key actions to be considered for a range of core digital preservation functions. [21] The purpose of this chart is to split tasks associated with digital preservation into manageable steps.

The Levels of Digital Preservation recommendations were developed by the National Digital Stewardship Alliance (NDSA), formerly a unit of the Library of Congress but now part of the Council on Library and Information Resources. They have further been adapted by the US Geological Survey as the "USGS Guidelines for the Preservation of Digital Scientific Data.[22]" However, the table below represents the original NDSA steps for each level. The NDSA also maintains a number of other recommendations documents which may prove useful to NAL moving forward, such as "Checking your Digital Content" which outlines the major issues around file fixity in digital preservation and ways to manage it, including checksums and scheduled fixity checks.[23]

[21] Phillips, Megan, et al. "The NDSA Levels of Digital Preservation: Explanation and Uses." Archiving Conference. Vol. 2013. No. 1. Society for Imaging Science and Technology, 2013.
[22] http://ndsa.org/documents/USGS_Guidelines_for_the_Preservation_of_Digital_Scientific_Data_Final.pdf
[23] http://hdl.loc.gov/loc.gdc/lcpub.2013655117.1

NSDA Levels of Digital Preservation

	Level 1: Protect your data	Level 2: Know your data	Level 3: Monitor your data	Leve 4: Repair your data
Storage and Geographic Location	Two complete copies that are not collocated For data on heterogeneous media (optical discs, hard drives, etc.) get the content off the medium and into your storage system	At least three complete copies At least one copy in a different geographic location Document your storage system(s) and storage media and what you need to use them	At least one copy in a geographic location with a different disaster threat Obsolescence monitoring process for your storage system(s) and media	At least three copies in geographic locations with different disaster threats Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems
File Fixity and Data Integrity	Check file fixity on ingest if it has been provided with the content Create fixity info if it wasn't provided with the content	Check fixity on all ingests Use write-blockers when working with original media Virus-check high risk content	Check fixity of content at fixed intervals Maintain logs of fixity info; supply audit on demand Ability to detect corrupt data Virus-check all content	Check fixity of all content in response to specific events or activities Ability to replace/repair corrupted data Ensure no one person has write access to all copies
Information Security	Identify who has read, write, move and delete authorization to individual files Restrict who has those authorizations to individual files	Document access restrictions for content	Maintain logs of who performed what actions on files, including deletions and preservation actions	Perform audit of logs
Metadata	Inventory of content and its storage location Ensure backup and non-collocation of inventory	Store administrative metadata Store transformative metadata and log events	Store standard technical and descriptive metadata	Store standard preservation metadata
File Formats	When you can give input into the creation of digital files encourage use of a limited set of known open formats and codecs	Inventory of file formats in use	Monitor file format obsolescence issues	Perform format migrations, emulation and similar activities as needed