

Digital Workflows at the National Agricultural Library and Implications for Preservation

February 2018

Morgan Daniels
Postdoctoral Fellow, University of Maryland College Park
morgan.g.daniels@gmail.com

Executive Summary

This study was designed to surface needs for an organization-wide digital preservation infrastructure at the National Agricultural Library by examining the processes currently used at NAL in routine work with digital materials. It used an observation-based interview method to learn directly from staff members about their workflows with digital objects, combining the information gathered into models that depict their work. The report is organized to follow each of the four major digital workflows, ending with a discussion of the implications of the study for an overarching digital preservation program at the library.

The discussion of digital workflows begins with the process of digitizing materials from the collection (page 7). Digitization workflows and their implications for preservation are aligned with the group doing the digitization: when digitization is done in-house by the Digitization and Access branch, preservation copies of documents are saved to an NAL drive, but when Internet Archive contractors complete the process, materials are stored on Internet Archive servers and NAL does not host a preservation copy. This report strongly suggests building an automated process into this workflow to obtain preservation copies of Internet Archive hosted materials for NAL. In addition, some ad-hoc digitization is done throughout the library to serve specific needs (particularly patron requests for portions of documents) and in these cases, the resulting digital documents should not be considered preservation-worthy. The second digital workflow examined in this report is the creation of online exhibits (page 13). Online exhibits and NAL created and collected websites should be considered candidates for robust digital preservation, with an emphasis on the content and format of the communication. The underlying digital materials used in exhibits should be preserved primarily through the digitization workflow, which creates a high quality digital copy. The third process, metadata quality control (page 19), is situated largely within NAL's Unified Repository. Issues in this workflow are related to system response times and a need for more staffing, to keep up with increasing demands. Finally, the report explores the process of curating research data (page 23) in which staff use several processes to complete similar work in several systems housing different types of data. While these systems offer a valuable range of interactions for data users, this report recommends designating Ag Data Commons as a single preservation repository for all data types hosted by NAL.

While this research suggests improvements to each of the workflows discussed in the report, there are several larger takeaways that affect NAL including, but also beyond, the digital practices discussed here. Almost every interviewee mentioned understaffing as a debilitating factor at NAL. A larger staff would ease the burden of current employees, who in one group reported that they complete the workload that was once done by a team of ten, with a current team of only two people. Understaffing causes employee burnout, process delays, and an inability to focus on innovation, as smaller groups of people attempt to meet or exceed productivity levels of a larger staff in the past.

The need for a comprehensive digital preservation program at NAL is both a motivating factor and a major finding of this report. The sections of the report detailing consolidated workflow models offer specific recommendations for improving workflows, while the proposed preservation system diagram and discussion at the end of the report give suggestions for improving the digital preservation practices of NAL as a whole. In brief, these suggestions are to update the Fedora installation used by the Unified Repository to take advantage of the robust preservation features offered in newer versions of the software and accommodate a greater range of digital materials and to join a digital preservation consortium to take advantage of the greater scale of preservation infrastructure that such consortia make possible.

A third organization wide need surfaced by this report is the need to communicate about workflows and practices between and among units. Improved communication about practices will help employees make connections between their work and their colleagues', benefitting cross-unit collaboration on shared concerns. Several initiatives at NAL push the organization in the right direction toward this goal, including the Brown Bag series, where staff present to each other about their work; and cross-organizational working groups, such as one which recently met about metadata types, workflows, and standards across the library. These and similar efforts should be supported, as they contribute greatly to the culture at NAL.

Digital preservation itself provides an opportunity for cross-organizational collaboration because it effects the work of all branches of NAL. Such a collaboration is suggested in the conclusion to this report, to be structured as a cross-organizational digital preservation working group, with leadership from ISD and participation from two staff members from each branch. For each unit, beginning with the findings around current practices in each group as discussed in this report, the staff members from that unit should review and verify the findings given here, augmenting them with their own findings about materials that require preservation and key workflow steps, if missing from these models. As a group, this team would determine the retention and preservation requirements of the many types of digital materials, resulting in an ever-growing shared register of digital assets at NAL. More suggestions for this collaborative group can be found in final section of this report, which looks at digital preservation from a library-wide perspective.

ISD is a good choice to spearhead this project, due to their current role in backing up library assets. At present, NAL servers and virtual machines are backed up daily and written to tape (which is stored offsite) weekly, with an annual back up copy also stored offsite. The current back-up practices are a necessary, but not sufficient, step towards robust digital preservation. The gap between current back-up practices and a robust digital preservation plan, and ways to bridge it which surfaced through the course of this study, are discussed throughout the report.

Table of Contents

Executive Summary	2
Study Design	5
Participants	5
Table 1. Interview Participants	5
Workflows	6
Table 2. Workflows Documented and Consolidated	6
Work Process 1: Digitizing collection materials	7
Table 3. Participants interviewed for work process 1: Digitizing collection materials	7
Digitization Scenarios	7
Consolidated workflow: Digitizing materials from the collection	8
Analysis and Recommendations: Digitizing collection materials	11
Digitization standards	11
Work Process 2: Creating an online exhibit	13
Table 4. Participants interviewed for work process 2: Creating an online exhibit	13
Consolidated workflow: Creating an online exhibit	14
Analysis and Recommendations: Creating an online exhibit	16
Work Process 3: Metadata quality control for publications	19
Table 5. Participants interviewed for work process 3: Metadata quality control for publications	19
Consolidated Workflow: Metadata quality control for publications	20
Analysis and Recommendations: Metadata quality control for publications	22
Work Process 4: Curating research data	23
Table 6. Participants interviewed for work process 4: Curating research data	23
Consolidated workflow: Curating research data	24
Analysis and Recommendations: Curating research data	28
Conclusion: Recommendations and system diagram for digital preservation at the library as a whole	30
Infrastructure Recommendations	30
Workflow Recommendations	30
Proposed Digital Preservation System Diagram for NAL	32
Organizational Recommendations	32
Appendix 1 Project Charter	34
Timeline	35
Responsibilities	35
About Contextual Inquiry	35
Appendix 2 Interview Protocol	36
References	37

Study Design

This study was conducted using the Contextual Inquiry method, developed and described by Karen Holtzblatt and Hugh Beyer. Specifically, it used methods enumerated in Holtzblatt, Wendell, and Wood (2005). Originally created in the context of user experience design, Contextual Inquiry helps researchers capture and understand the work participants do in a particular organization or work setting. Contextual Inquiry interviews typically begin with a brief set of questions (see appendix 2 for the semi-structured interview protocol used in this study) followed by an in-depth walk-through of the processes individuals use to do the work under study. The walk-through requires investigators to adopt an apprentice role, learning about work processes by observing and asking questions about current or recent concrete instances of the participants' work.

The scope of this study concerned the creation and management of digital objects throughout the National Agricultural Library, seeking to discover the ways in which staff members process, describe, alter, store, and preserve digital objects. At the conclusion of each interview, the researcher developed a visual model of each workflow discussed. By consolidating related models into one overall model for each process, the researcher produced broad scale depictions of the work required to produce a digital product, providing an organization-wide perspective on work with digital objects. Through these consolidations, presented and described in this report, NAL staff and leadership can better understand how the work of many units contributes to the library's overall efforts, while easily seeing problems that arise in the process. With these models, the digital preservation needs of NAL are more easily mapped and identified. See Appendix 1, the Project Charter, for more information about the study methods.

Participants

Twenty-two NAL staff members were interviewed for this study between February and August 2017- approximately one quarter of the Library's full time staff. Participants' length of tenure at NAL ranged from one year to 34 years, averaging 12 years across all interviewees. Beginning with the NAL staff list, the investigator emailed interview requests to individuals known to be involved with the creation, curation, or other management of digital objects. This approach was combined with a snowball sampling method using responses to question 9 of the interview protocol "Who else should I talk to about these issues, specifically people in your unit or people involved in your workflow for these materials?" (See Appendix 2 for the full interview protocol). As shown in Table 1, below, participants were drawn from all divisions of the National Agricultural Library, except for the Office of the Director, which does not have hands-on involvement with the workflows under investigation.

Table 1. Interview Participants

NAL Unit	Participants
DPD: Acquisitions and Metadata	3
DPD: Digitization and Access	4
DPD: Indexing and Informatics	1

IPD: Digital Library	2
IPD: Information Centers	3
IPD: Customer Services	1
ISD: Applications and Systems Technology Branches	3
KSD	5
Total	22

Workflows

There are four major digital object workflows discussed in this report, as seen in Table 2 below. These workflows represent the bulk of digital object processing that takes place at NAL, which is largely focused in the following areas: digitization of materials from the library’s collection; digital organization and communication of those and other materials through online exhibits; curation of research data; and metadata clean up and correction for resources created outside NAL, particularly agricultural science literature produced both within and outside USDA. These workflows emerged from discussion with staff members across NAL, through their responses to the second interview question “Please describe the types of digital materials you deal with in your work at NAL (including “born digital” materials, digital representations of analog material, and web hosted databases).” This purpose of this question was primarily to surface the range of digital materials and related workflows that were relevant to the interviewees’ daily tasks.

The majority of staff members interviewed for this study participated in the four workflows addressed here, while several individuals were involved in more than one of them. The number of participants listed in Table 2, therefore, does not directly match the number of participants in this study. The following sections of this report discuss each workflow in turn.

Table 2. Workflows Documented and Consolidated

Workflows	Participants
Digitizing materials from the collection	8
Creating an online exhibit	5
Metadata quality control for publications	6
Curating research data	5
Total	24*

*several interviewees discussed multiple workflows

Work Process 1: Digitizing collection materials

The process of digitizing materials for the NAL collection was a clear choice for inclusion in this study. With the involvement of numerous people throughout several units of NAL and the Internet Archive, all working together to convert print-based collections into digital files, it presents an interesting organizational and workflow challenge. Of the 22 people interviewed for this study, 8 were directly involved in the process of digitizing materials from the collection, in roles ranging from selection of objects for digitization to managing the overall digitization process. Their organizational affiliations within NAL are represented in the following table.

Table 3. Participants interviewed for work process 1: Digitizing collection materials

NAL Unit	Process	Number of Interviewees
DPD: Digitization and Access	Digitizing collection materials	2
DPD: Digitization and Access	Managing digitization workflow (both NAL and Internet Archive)	2
IPD: Information Center	Finding materials for a digital collection	2
IPD: Customer Service	Digitizing materials in response to a user request	1
DPD: Acquisitions and metadata	Cataloging digital collections	1
Total		8

Digitization Scenarios

Three distinct digitization scenarios emerged from interviews with staff, corresponding with different processes used at NAL. The Digitization and Access branch manages both their own digitization process (shown in green) and the work done by the Internet Archive (in yellow), which contracts much of the digitization work at NAL. Digitization begun in response to a user request (in blue in the model below) might take place entirely outside of the Digitization and Access branch of the library, done instead by the Customer Services branch staff member receiving the request.

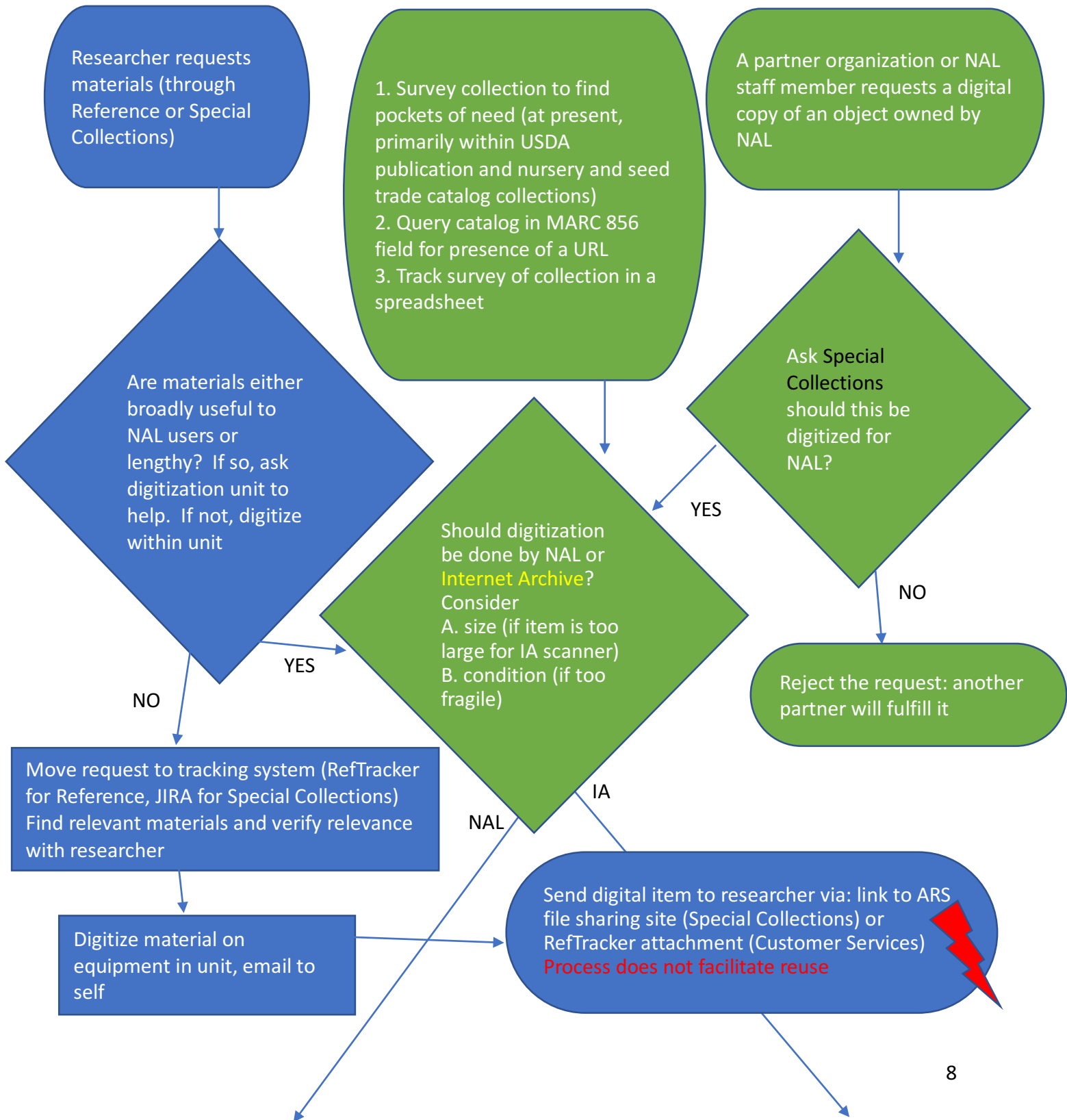
Note: Other NAL units involved in the model below are shown in black type, while yellow type indicates a group outside NAL. A red lightning bolt icon indicates a breakdown in a process.

Consolidated workflow: Digitizing materials from the collection

Scenario 1: Workflow is triggered by a researcher request

Scenario 2: Workflow is triggered by the process of systematic digitization using Internet Archive

Scenario 3: Workflow is triggered by a request from a partner organization



Digitization fulfilled by
Digitization & Access group at
NAL

1. Create an entry in Jira, assign to staff member: track hours spent on process for an idea of labor costs
2. Mark the request "In progress" in partner's admin tracking module (IBHL's tool, Gemini)

Software creates both TIFF and JPG2000 derivatives, moving TIFF to Special Collections Masters folder as the preservation copy
Access copy, JPG2000, is moved to Ready for Uploading folder, which also contains a MARC xml file for the serial record, exported from OCLC

1. Assigned staff member retrieves items from stacks, leaves a shelf marker, brings items to digitization area
2. Staff member verifies catalog record, enters MARC 016 and call number in a csv, and notes Special Collections approval
3. Logs work in Jira as volume scanning is begun and completed
4. [if applicable] Adds partner-required metadata for each volume

Digitization fulfilled by Internet
Archive

- Once selected, pull all copies from shelf
1. Sort and collate, selecting best copies (least fragile for Internet Archive to handle)
 2. Remove staples, paperclips if they get in the way of scanning process
 3. Create a Jira record for each batch

1. Use Perl script from ISD Applications Branch to pull metadata from catalog and format for IA in XLS
2. If there are errors in the catalog records, make a spreadsheet detailing typos, title matching, other issues, and send monthly to the cataloging group email address

Internet Archive staff return a spreadsheet which itemizes the number of fold outs and image counts, and gives links to published images in IA

- Quality Review procedures instruct staff to review a sample of the returned materials, looking at thumbnails and metadata
1. Does title match? Is image good? Is disclaimer page present?
 2. If there are missing pages, also check neighboring items
 3. Create a worksheet within the spreadsheet to track required changes
 4. Send xls to Internet Archive site manager for correction
 5. Once corrected, alert IA that NAL has accepted the batch

Send spreadsheet to **ISD Applications Branch** on a monthly basis to add Internet Archive links back to NAL catalog in 856 field-- Location: Electronic Resource
Update spreadsheet tracking digitization progress in the collection
Materials are not systematically harvested from Internet Archive- NAL does not own or preserve a digital copy

A regular Cron job (Perl script) from **ISD Applications Branch** uploads the materials to **Internet Archive**. It runs at noon and 4pm to see if any new materials need upload. Sends an email stating successful or failed **Uploading can be a problem if a batch is too big, easier to troubleshoot with a smaller quantity**

Analysis and Recommendations: Digitizing collection materials

The digitization process is the most complex workflow consolidation presented in this report. Because digitization requests may arise from several sources, including NAL users, partner organizations (such as the Biodiversity Heritage Library), and NAL staff member requests, procedures vary somewhat depending on origin of the request and the purpose to which the digital object will be put. The diagram above consolidates all three of these scenarios into one model, using color to differentiate between them. Yellow is used to depict digitization done by the Internet Archive as an NAL contractor, green depicts the process as fulfilled by Digitization & Access branch members, and blue shows digitization done within another unit of the library.

Each of the digitization scenarios has implications for the preservation and reuse of digital objects. Both the green and yellow diagrams above show that NAL uses the Internet Archive as a preservation platform, creating an NAL-owned version of a digital document on an NAL storage drive labeled “Masters” only when materials are digitized in-house. In the yellow, Internet Archive digitized condition, we see that from the perspective of NAL’s digital preservation responsibilities, the process is inadequate. NAL staff add Internet Archive links to their catalog, but do not systematically download the files to the NAL infrastructure. One major recommendation of this report is the addition of a process to routinely transfer new files created by the Internet Archive into NAL’s Unified Repository. Master copies of these digital files (derived from both workflows) should be part of a robust NAL-controlled digital preservation infrastructure for long-term stability.

In the blue diagram, depicting cases in which a staff member in a customer services role decides to digitize some part of an item in response to a user request, the document is essentially lost immediately from a preservation and reuse perspective—it has only short-term usefulness, in answer to a request, and will not be saved in a way that facilitates future finding or reuse. (Alternately, if the document is deemed to have long term usefulness, it joins the Digitization and Access team’s workflow, making it more broadly accessible.) Several staff members in the special collections and customer service groups voiced the desire to make these materials available for reuse. This report recommends, however, that because such materials are digitized in an ad hoc way, without conforming to shared quality or storage standards, and often do not represent an entire publication but an excerpt instead, they should continue to be considered single-use copies, to be retained only for the immediate future but not the long term.

Digitization standards

As a source for technical standards for digitization, the Digitization and Access branch of NAL follows the Federal Agencies Digital Guidelines Initiative (FADGI) Still Image Working Group’s publication Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster

Image Files (Rieger 2016). This document provides recommended standards for digitizing still image materials, including factors like master file formats, resolution, and bit depth. For each material type, the FADGI guidelines give recommendations using a one-to-four star rating system, where one star represents low quality images and four stars represent best practices. “Three star imaging defines a very good professional image capable of serving for almost all uses. Four star defines the best imaging practical today. Images created to a four star level represent the state of the art in image capture and are suitable for almost any use” (Rieger 2016, p 9). By aligning imaging standards with FADGI guidelines for three and four star ratings, the NAL Digitization and Access branch assures that the images created in the digitization process will be high quality and adaptable to a myriad of end-user needs. These standards should continue to be monitored and adopted by NAL, and adhered to by digitization contractors. By assuring that FADGI digitization standards are followed, NAL brings itself into line with standards followed by other federal agencies while creating the highest quality digital surrogates possible, appropriate for long term preservation.

Work Process 2: Creating an online exhibit

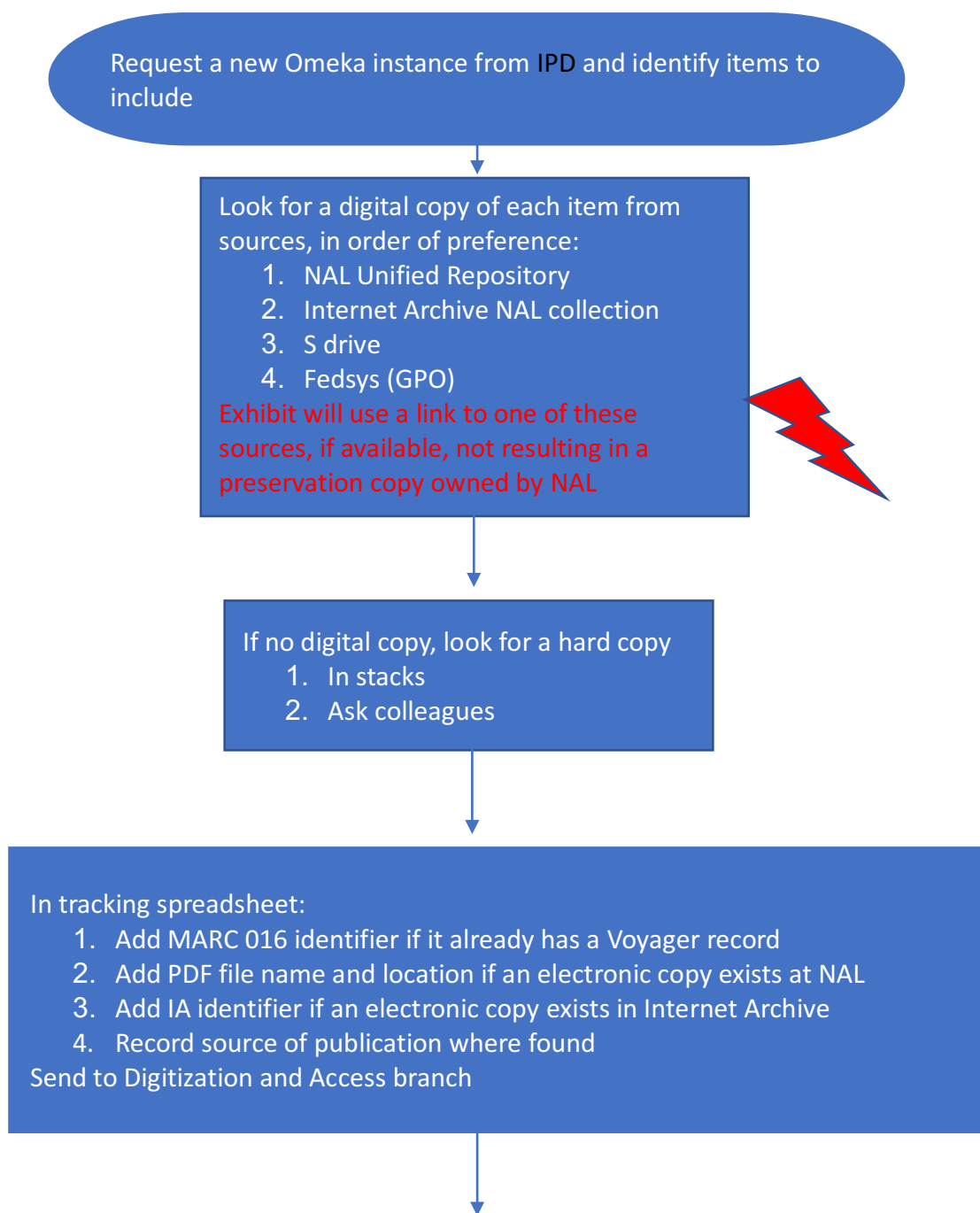
The second workflow presented in this study is the creation of online exhibits using the Omeka web publishing platform. Omeka is used largely by two groups within NAL: the Information Products Division (IPD) and the Special Collections unit. Of the 22 people interviewed for this study, 5 were directly involved in the process of creating Omeka exhibits for the library. Their organizational affiliations within NAL are represented in the following table.

Table 4. Participants interviewed for work process 2: Creating an online exhibit

NAL Unit	Process	Number of Interviewees
DPD: Digitization and Access	Creating an online exhibit	1
IPD: Information Centers	Processing materials for a digital collection	3
IPD: Digital Library	Creating an online exhibit	1
Total		5

In the following diagram, the color blue represents actions taken primarily by members of the Information Products Division or the Special Collections group, as they seek to create a new online exhibit. Green is used to depict the work done by the Digitization and Access group to digitize materials for a web exhibit. As in the previous workflow diagram, other NAL units involved in the model below are shown in black type, while yellow type indicates a group outside NAL. Red lightning bolt icons indicate breakdowns in a process.

Consolidated workflow: Creating an online exhibit



Digitization and Access adds columns for:

1. Digitization flag number,
2. Next Steps: "Harvest from IA"; "Tag for (Info Center)"; "Add collection name to catalog"; "Upload to UR, MODS to MARC transformation" for items found in FedSys.

Digitizes item (where needed)

Forwards the spreadsheet to ISD to add persistent identifier (agid)

Digitization team returns an Excel spreadsheet with links to digitized versions of publications in [Internet Archive](#)

Retrieve each item via the Internet Archive link to download selected images to desktop

Open Omeka to add digital item/s, supply Dublin Core metadata, copy and upload an excerpt from the text, where appropriate



Use Omeka plug in for exhibits to link and organize items



Create a timeline using timeline.js if appropriate to the exhibit



Send Word version of the exhibit to the **USDA Office of Communication** for editing/ approval



Once approved, IPD runs a 508 compliance check using Total Validator Pro



Test the site and when ready, ask ISD to move it to the Production server, where it is live and publicly available **No robust web preservation practices are currently in place. While servers are backed up by ISD, web preservation is not yet done at NAL**



Analysis and Recommendations: Creating an online exhibit

In the course of interviews with NAL staff, it emerged that IPD and Special Collections currently use different Omeka templates for their sites, resulting in exhibits with two distinctive looks. Examples of current sites created by the two groups follow:



How Did We Can? The Evolution of Home Canning Practices
<https://www.nal.usda.gov/exhibits/ipd/canning/>



An Illustrated Expedition of North America
<https://www.nal.usda.gov/exhibits/speccoll/exhibits/show/an-illustrated-expedition>

In practice, all Special Collections Omeka materials are grouped together as “exhibits,” in “one big bucket” as one staff member described it, while IPD exhibits are each stand-alone Omeka instances. As a consequence, IPD exhibits can each have a unique look and feel, which is appreciated by IPD staff, who feel that this adds visual interest to the exhibits. The Special Collections group’s approach puts a greater emphasis on standardization, representing each exhibit as part of a Special Collections Exhibits product.

Having multiple Omeka instances has greater implications for upkeep and maintenance than it does for digital preservation. ISD must be asked to update multiple sites when a new version of Omeka is adopted by NAL, which may be more difficult the more variation is used in the exhibits. This workflow and standardization question should be addressed through a conversation between ISD, IPD, and Special Collections stakeholders, who may decide that the value to NAL of flexibility of individual Omeka instances outweighs the difficulty of updating multiple instances.

From a digital preservation standpoint, however, this question does not loom large. Web exhibits pose two kinds of preservation challenges: the preservation of the underlying content used in an exhibit and the preservation of its packaging as an online exhibit. The workflows used by NAL staff to create web exhibits suggest an emphasis on the latter type of preservation rather than the former. Since exhibit creators use *excerpts* of materials found elsewhere at NAL or on the open web, the goal in preserving web exhibits should be in preserving the content of NAL communication with the public, *not* preserving the underlying digital objects, which should be maintained through preservation of the systems on which they reside as complete documents (i.e. the Unified Repository).

From a digital preservation standpoint, current NAL back-up measures protect a copy of digital exhibits, but a more comprehensive approach to web preservation is in order, to capture and package content for long-term preservation. The scope of NAL websites to be preserved is broader than web exhibits, including information centers, data products, the library's home page, and other NAL provided content (along with other web materials NAL may choose to collect). Emphasis on the Omeka based exhibits in this report is useful, however, because it illustrates the variation in use of a single tool in departments throughout the library.

NAL should consider two products from the Internet Archive for preserving web content. Archive-It is a hosted service widely used in research libraries while Heritrix is an open-source solution, which would require more hands-on work by NAL staff in its implementation (while some might consider the open-source option "free," the cost to already taxed ISD staff time is in no way negligible).

Work Process 3: Metadata quality control for publications

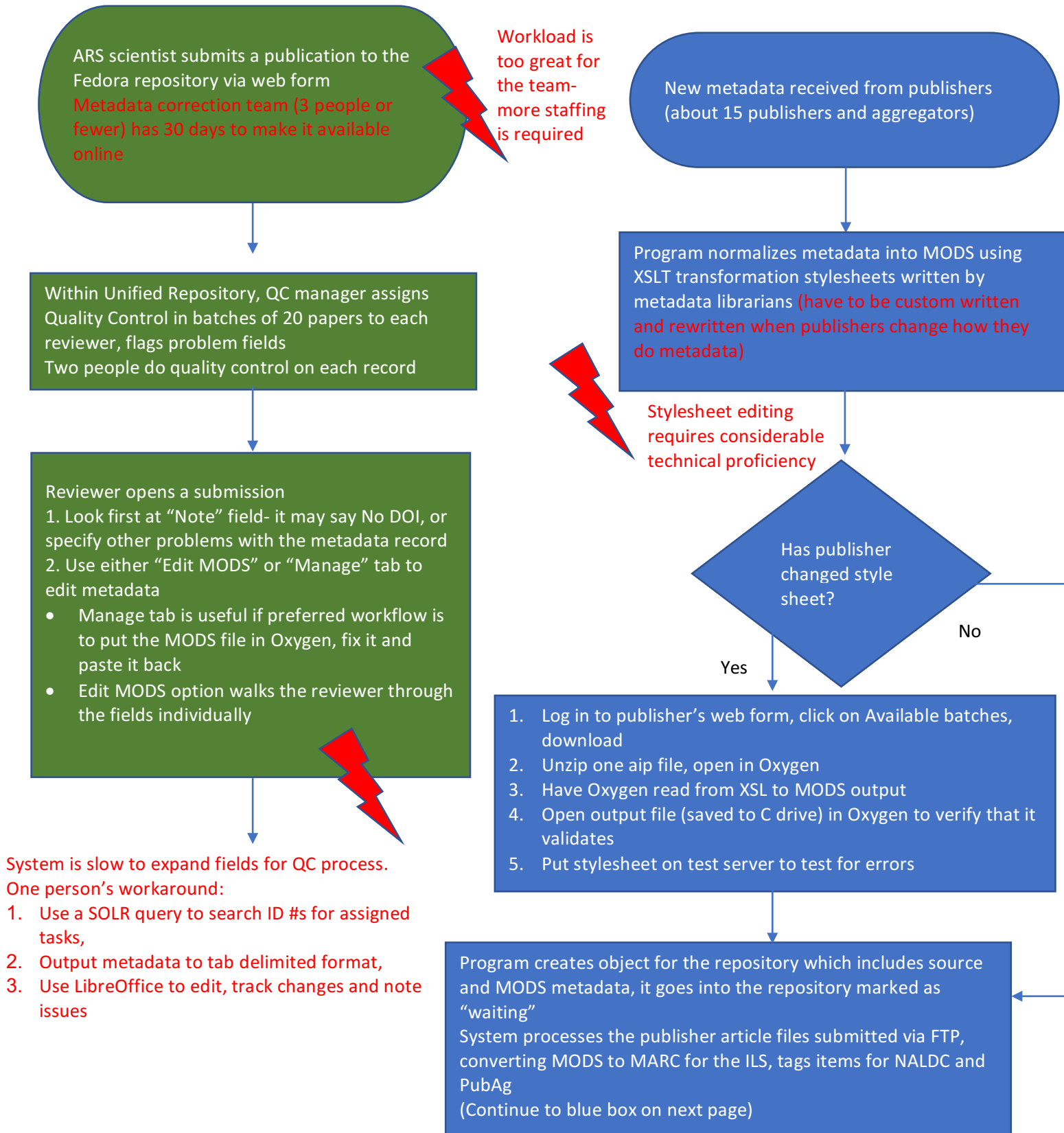
Metadata quality control at NAL takes place in many different contexts, with numerous types of objects. Metadata quality control for publications, specifically, is a key process involving the work of several people invested heavily in converting metadata from publisher feeds and author provided sources to formats compatible with NAL systems. Of the 22 people interviewed for this study, 6 were directly involved in ongoing metadata cleanup. Their organizational affiliations within NAL are represented in the following table.

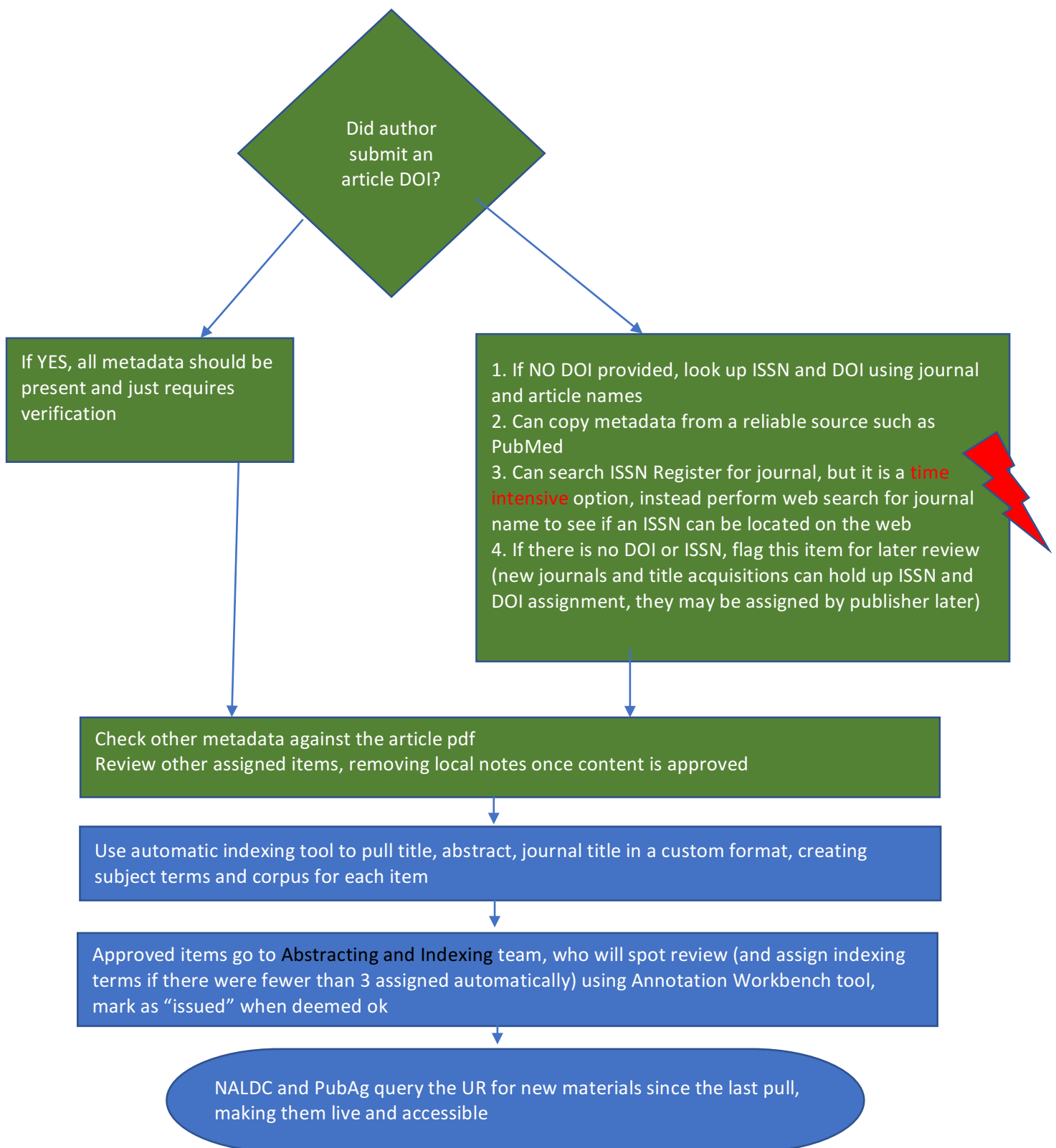
Table 5. Participants interviewed for work process 3: Metadata quality control for publications

NAL Unit	Process	Number of Interviewees
DPD: Indexing and Informatics	Subject indexing of journal articles	1
IPD: Digital Library	Manage review process for USDA author submitted articles	1
ISD	Correct conversion issues in publisher-provided metadata in MARC to MODS transformation	1
ISD: Systems Technology	Create article information for PubAg	1
DPD: Acquisitions and metadata	Metadata quality control	2
Total		6

Note: Two primary workflows emerged in the process of metadata quality control, which are both depicted in the following model. The two workflows are characterized by the metadata source—in green, articles submitted by ARS scientists for inclusion in the NAL repository, and in blue, metadata feeds received from publishers. The final steps, abstracting and indexing and publication (shown in dark blue), apply to both workflows. Red lightning bolt icons indicate breakdowns in a process.

Consolidated Workflow: Metadata quality control for publications





Analysis and Recommendations: Metadata quality control for publications

The processes used at NAL for metadata quality control are largely mechanized, and interviews surfaced few complaints with the processes themselves. Where issues did arise for staff members, they related to slow response times in the Unified Repository, time intensive aspects of the work (such as searching the ISSN Register), and understaffing issues that made the work more difficult. While systems development at NAL has served metadata quality control processes well, there is clearly a need for a larger team of individuals to work on this process. Some aspects of metadata quality control (particularly working with publisher stylesheets and metadata transformation) require specialized technical proficiency, reducing the available pool of NAL staffers who might currently be able to join this team. However, reskilling and professional development of current metadata staff would help build more capacity for this group. In addition, hiring new staff who can jump into the metadata correction process should be a priority for the Library.

When staffers found problems with the workflows in place, they developed their own workarounds. One example is an individual's use of SOLR to query the UR and perform work in LibreOffice, to avoid slow reaction times in the QC nodes of the Unified Repository. This workaround signals room for improvement in the UR infrastructure—quicker response time would make the process simpler for users. On the other hand, only one of the six staff members working with metadata quality control mentioned slow system times as a problem, suggesting that the others did not perceive system times as an issue.

Because metadata correction workflow is primarily mediated through the Unified Repository, the UR is a major target for digital preservation of this content. With robust preservation of the UR in place, these materials should be secured for long-term access and use. Steps to increase the preservation capabilities of the UR are discussed at the conclusion to this report.

Work Process 4: Curating research data

At the National Agricultural Library, research data curation takes place within the Knowledge Services Division, but work processes vary greatly in that division depending on the systems used for each product, as the model below illustrates. Of the 22 people interviewed for this study, 5 were directly involved in research data curation. Their organizational affiliations within NAL (all within the Knowledge Services Division) are represented in the following table.

Table 6. Participants interviewed for work process 4: Curating research data

NAL Unit	Process	Number of Interviewees
KSD: Scientific Data Management	Curating research data	2
KSD	Curating research data	3
Total		5

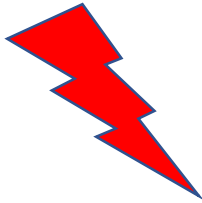
Because of the variation in work taking place between the three data systems studied for this report (Ag Data Commons, Lifecycle Assessment Commons, and i5k) the three systems are depicted independently in this model. Processes can be read in parallel, with a summary of each step in a blue figure at the top of each page.

Note: Other NAL units involved in the model below are shown in black type, while yellow type indicates a group outside NAL. Red lightning bolt icons indicate breakdowns in a process.

Consolidated workflow: Curating research data

Researcher submits data/ model/ genome, which generates notification for curator

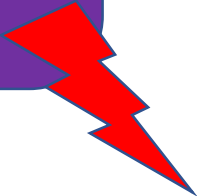
Open and view files locally
Each data product is in a separate system, requiring upkeep of multiple tools



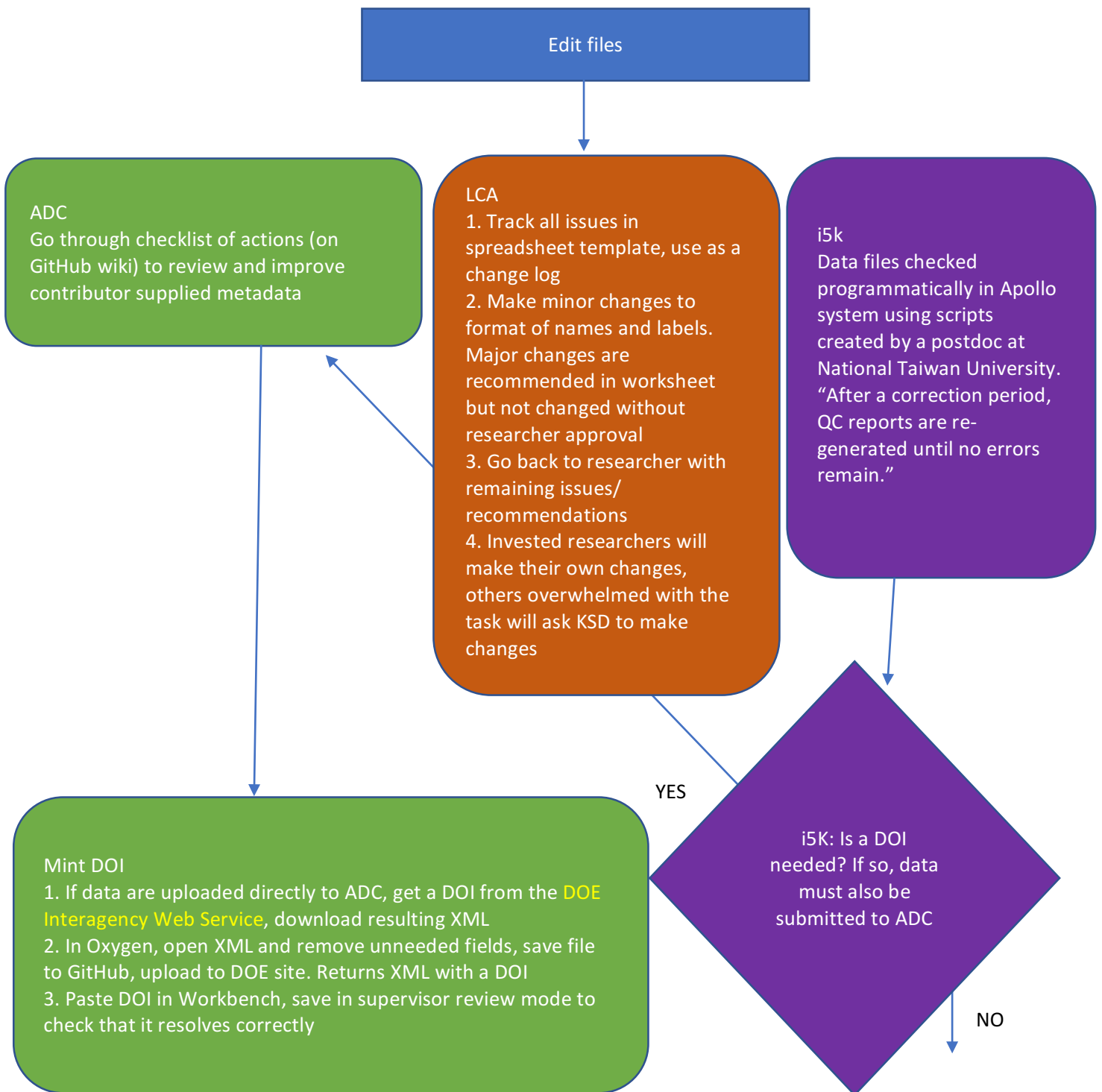
Ag Data Commons (ADC)
1. Log into Ag Data Commons Workbench (back end interface)
2. Review record for any apparent errors

Life Cycle Assessment Commons (LCA)
Check fidelity of model:
1. Create a new database, import files to see that they open properly, close database
2. Open SQL client, run SQL scripts that check models run by specified rules
3. Read and verify documentation

i5k
1. View submitted files and information (Submission to NCBI is a prerequisite, allowing i5k to benefit from the NCBI ingest process)
2. Run md5sum checksum to verify that transfer happens correctly (there are often issues)




Checksum is routinely run only in i5k, the other systems should also run similar checks



Review



ADC

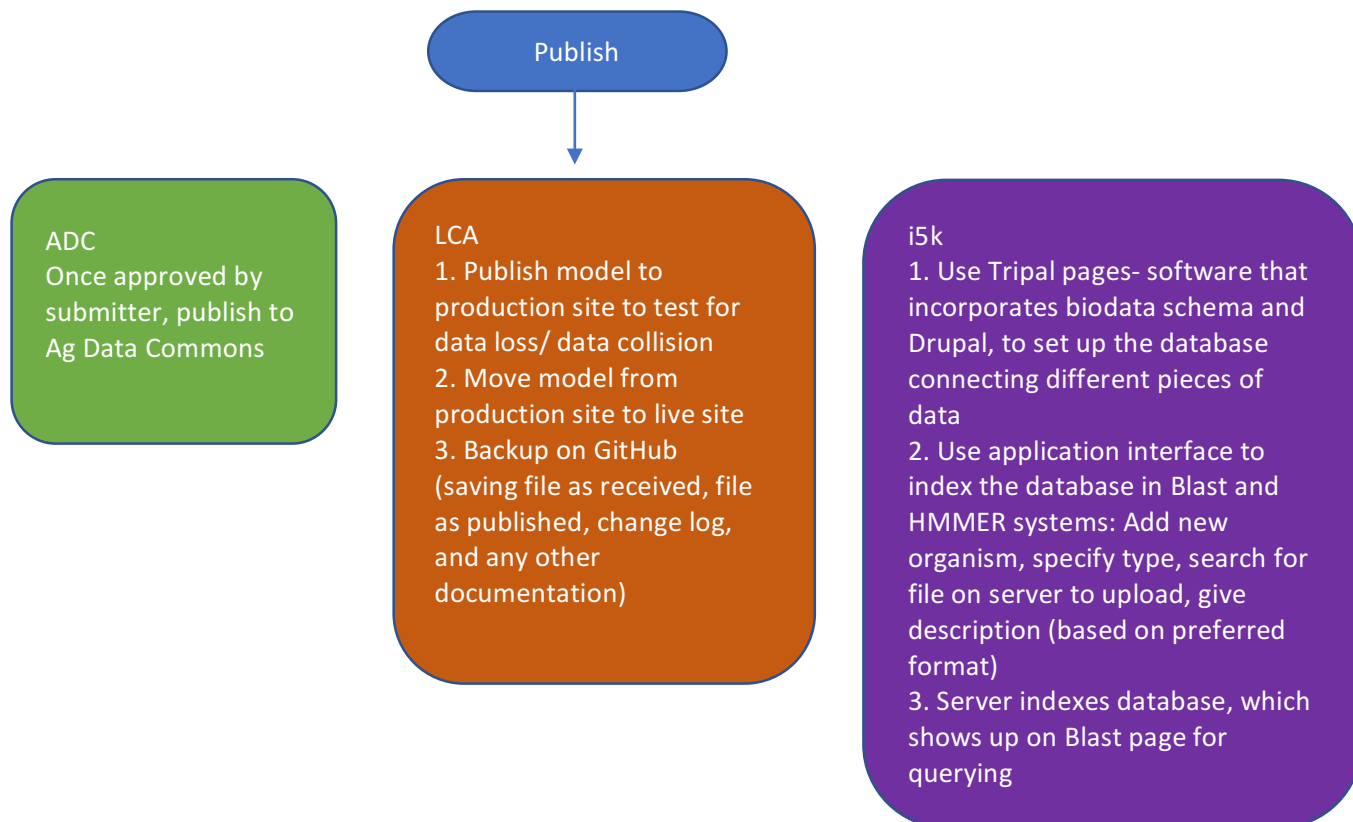
1. Email list of changes to author for approval and ask any questions
 2. Authors resubmit data for review. If they change anything: they email with a list of changes rather than using a modification note, **but curators worry this method won't scale up when data submissions increase**
- 

LCA

1. **Independent peer reviewer** with subject expertise reviews documentation and representation of model, making comments (reviewers are network partners, some academic, some government)
2. Curator reviews comments from reviewers, but will only relay comments related to documentation back to the researcher
3. Researcher revises if necessary (may ask KSD to help)

i5k

Email discussion with depositor about any unclear metadata or other issues



Analysis and Recommendations: Curating research data

The most striking element of this workflow model is the variation in processes between systems for working with data. Differences in funding, partnership models, and project histories account for many of the differences in systems and workflows. The i5k project, for example, is a community effort of arthropod researchers for whom a shared database of sequenced genomes of these animal species is intended to improve research outcomes (Poelchau 2015). In contrast, the Life Cycle Assessment Commons supports assessment of “environmental impacts associated with all stages of a product's life. [...] The goal of LCA is to compare the full range of environmental effects assignable to products and services by quantifying all inputs and outputs of material flows, and then assessing how these material flows impact the environment” (Life Cycle Assessment Commons). In addition to raw data related to the assessment of numerous agricultural products, the LCA Commons contains the models used by researchers to assess the impact of actions throughout an agricultural product’s lifecycle. Between i5k and LCA alone, there are a host of differences in data structures, infrastructure requirements, and quality control methods. Ag Data Commons (ADC in the model above) contains a wider range of data types, seeking to serve as a catch-all data repository for agricultural researchers. What ADC gains in scope, it loses in the ability to standardize, which is an important consideration for the long-term usability of datasets. While omitted from the workflow model due to space constraints, LTAR, the Long-Term Agroecosystem Research (LTAR) network is a fourth data product provided by KSD, offering historical data from 18 research sites with an average of 50 years of data. As a repository for data over a long period, it is an extremely valuable resource. Because it focuses on observational data, geographic location is of central importance to LTAR, whereas location is less significant for the other data systems managed by KSD.

While separate systems for groups of research data provide unique forms of access to the data they contain, this does create some redundant work in system upkeep. To the extent possible, NAL should attempt to bring the systems together and develop features that can be accessed across platforms. For example, one major goal of KSD is to connect research data with the publications derived from that data. In Ag Data Commons, the metadata field “Primary Article” can accept a citation, DOI, and AgID, facilitating linkages between ADC, other NAL products, and resources found elsewhere on the internet. One possible workflow change would involve making an ADC record for each new dataset, which KSD might then treat as NAL’s preservation copy of the data and metadata. While the other data products would continue to offer enhanced functionality for using the data, ADC would serve as the umbrella repository for preservation purposes.

In the first stage of the model, receipt of files, curators for each of the three systems open and view the materials they have received, checking for issues in the integrity of the files. While i5k and LCA have well defined error checking methods (checksums and script running) ADC uses a more general review method, reflecting the broad range of data types the system accepts. Adding automated checksums to the submission process for all NAL data systems would help ensure that file transfer happens correctly, and increase the trustworthiness of these systems.

One notable shared feature of the three workflows is the documentation of the editing process. Whether done programmatically, by following a checklist, or with changes documented in a spreadsheet, KSD ensures the integrity of the data in part through keeping a record of transformations made. In the review section of the three workflows, only LCA uses peer review to ensure the quality of the submission. The other tools use back and forth communication with depositors to come to an agreement about changes that need to be made. This is a labor intensive method, as staff members working on ADC noted, and they are concerned that it may not scale up as submission volume increases. Offloading that work from staff to researchers through peer review is not necessarily the best option, since it demands sustained commitment to the resource from a community of expert users. As it grows, ADC may want to consider presenting data with different levels of curation, including self-deposited (no curation), minimal curation (curators have reviewed data and documentation for completeness), and peer reviewed (an expert reviewer has looked in-depth at the materials). Once a submission is ready for publication, curators use a range of manual and programmatic publication methods. ADC and LCA both have a curator move a data product into publication mode, while i5k uses indexing software to make data easier to query within the system's taxonomy.

One important aspect of data preservation, which KSD has configured in different ways for different data products, is documentation of data in its various states throughout curation. LCA uses a spreadsheet template to document changes, publishing it to GitHub along with the data as submitted, as published, and data documentation. In a sense, GitHub is LCA's preservation repository and the materials here are a core part of KSD's digital assets. The i5k and ADC servers are the primary targets for the digital preservation of those two products, although KSD staff should make careful note of where they store other materials and versions of data published to those systems—they likely also require long-term preservation.

Conclusion: Recommendations and system diagram for digital preservation at the library as a whole

This section of the report considers high level infrastructure and workflow changes that would support digital preservation at NAL, bringing together the findings of each section of the report with recommendations for building a more robust suite of systems.

Infrastructure Recommendations

The current state of digital storage and preservation at NAL can be described as system back-up, lacking a preservation focus. Through an automated process, a member of the ISD team oversees the daily backup of about 184 virtual machines at NAL, storing approximately 50TB of data. Using Veritas NetBackup Enterprise 8, NAL creates a disk-based snapshot of the virtual machines. On a weekly basis, the disk writes to tape, which is stored offsite by the vendor JK Moving. There is also a back-up copy of NAL data stored offsite and refreshed yearly. The technical gap between NAL's current digital storage back-ups and a robust preservation plan is primarily in methods to assure that the data will remain stable and accessible over time. Best practices for digital preservation suggest multiple copies of data on multiple servers in multiple geographic locations, with continuing automated fixity checks to ensure the authenticity of the data over time (Philips et al. 2013). While NAL uses two locations for data storage, they lack the recommended geographic distribution and automated fixity checks between copies that best practices recommend. A practical option to address this problem is membership for NAL in a group like the Digital Preservation Network or the LOCKSS Program (Lots of Copies Keep Stuff Safe, hosted at Stanford University Libraries), which provide robust systems for data ingest, replication, and fixity checks in exchange for a membership fee.

Some elements of current NAL infrastructure support digital preservation, and that capacity should be used to NAL's advantage. The Fedora software supporting NAL's Unified Repository is one prominent example. Fedora has several features supporting digital preservation that have been built into more recent versions of the software (Fedora 2018). By building a version of the Unified Repository using the 4.7.x release, NAL would be able to take advantage of the persistence, fixity, auditing, and versioning features of the software. The Unified Repository currently houses both in-house digitized materials, one portion of the materials produced in workflow 1 in this report, and the publication metadata produced in workflow 3. During NAL's Fedora installation upgrade, it should be considered as a potential in-house preservation system for Internet Archive produced materials, research data storage (as a central preservation solution for NAL's several systems), and archived NAL web content as well.

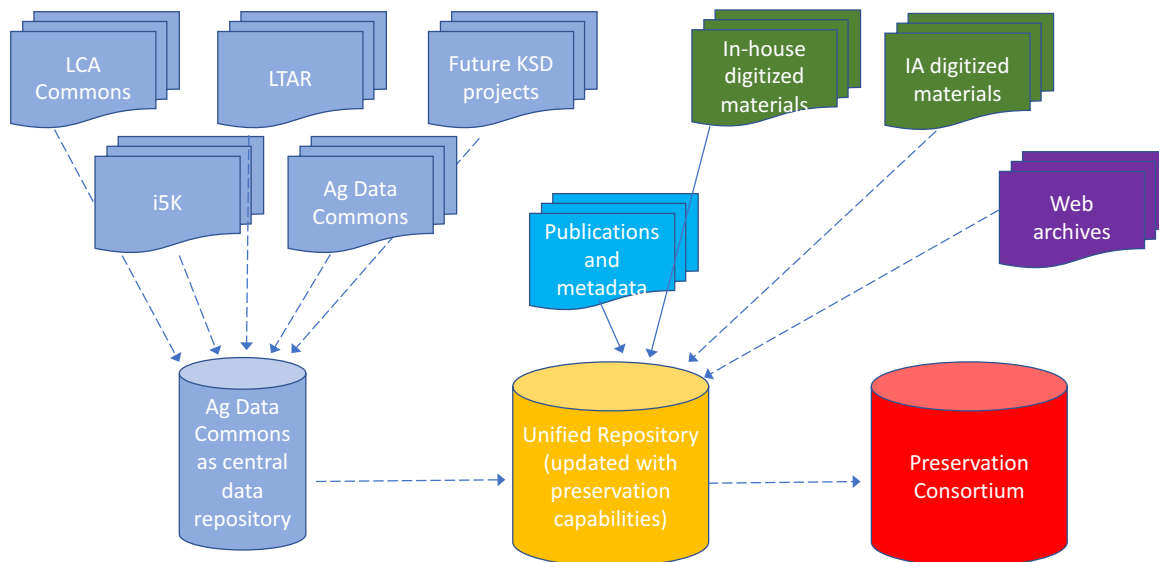
Workflow Recommendations

A key facet in the recommendations given in this report is an upgrade to a new Fedora installation. Although the Fedora software is open source and has no licensing fee, the upgrade would require significant labor costs to NAL, including earmarking the time of already overtaxed

members of ISD. While NAL tends to use contractors for major technical infrastructure upgrades, the library should use caution in relying too much on contractor labor. Contractors will need to work closely with NAL staff members managing and using the current installation to ensure that it is configured to meet their work needs and support workflows from many areas of the library. In particular, this report recommends several changes to current workflow practices to enhance preservation within the Unified Repository, which a new installation should support.

1. Build a new Unified Repository on a 4.7.x release of Fedora, taking advantage of the software's preservation capabilities (described in the previous section). This would provide robust preservation assurances for collections already housed in the UR.
2. Programmatically add all new and preexisting NAL content which resides in the Internet Archive to the Unified Repository. The script currently in use to upload new in-house digitized materials to IA was written in collaboration between ISD and Internet Archive staff. A similar collaboration should take place to write a script that reverses the process, providing NAL with its own digital copies of Internet Archive produced digital objects and their metadata, through deposit in the UR.
3. Research data are not currently housed in the Unified Repository, but if the repository is re-envisioned as a preservation environment, this should change. Because KSD is working towards recognition and use of Ag Data Commons as a disciplinary data repository for the agricultural research community, it should begin that process by cataloging and storing datasets from other KSD data products into the ADC system. ADC can serve as the repository for all KSD products, with links back to datasets in systems like i5k and LTAR that support enhanced exploration and use of the data. An (ideally automated) process should then be used to copy data to the Unified Repository, to provide back up and preservation support for the materials held in Ag Data Commons.
4. Web archiving is not yet one of the services performed by NAL, but it should be considered an essential part of a robust digital preservation system. Using software like Archive-It or Heritrix, NAL could begin to package and make accessible snapshots of the Library's own websites, along with other websites determined necessary to documenting agriculture. Web materials managed by ARS and other groups within USDA are particularly important targets for web archiving by NAL. Access to web archiving software and a clear collecting scope for web materials will enable NAL to collect this valuable part of agricultural history. The Unified Repository should be configured to manage and store archived websites.
5. With the Unified Repository updated and configured to support robust digital preservation, NAL should begin membership in a digital preservation consortium, such as the Digital Preservation Network or LOCKSS, to ensure the long-term preservation of materials stored both in the UR and in other NAL systems.

The following diagram illustrates the proposed relationships between current and not-yet-existent systems at NAL. Note that solid lines in the figure represent data flows already in place, while dashed lines represent proposed relationships.



Proposed Digital Preservation System Diagram for NAL

Organizational Recommendations

NAL should establish a preservation working group to be led by ISD with participation from two staff members from each branch. For each unit, beginning with the findings around current practices as discussed in this report, the staff members from that unit should review and verify the findings given here, augmenting them with their knowledge of materials that require preservation and key workflow steps, if missing from these models. This effort should result in an inventory of digital collections: a spreadsheet that captures information on digital preservation needs at the collection level, rather than at the object level. The inventory process is described in the Digital Preservation Workflow Curriculum compiled by the Digital Preservation Network (2018). Reviewing this curriculum as a group would be a valuable, yet accessible, learning exercise for members of NAL’s preservation working group, helping them begin with a shared understanding of digital preservation and how to achieve it. As listed in Module 2 of the training on the topic of selection, fields captured in a digital collections inventory should include:

- “Collection title/s
- Location/s of content
 - On which server or network drive? On external hard drives, DVDs, or CDs? In what box? On which shelf? In which room?

- Agents responsible for creating the collection (e.g., donor, digital collections department)
- Agents responsible for curating the collection (e.g., archivist, digital preservation librarian)
- Content stream (e.g., born-digital, digitized)
- Format
- Number of files
- Size of collection (in bytes)
- Collection creation date/s, date of initial inventory, event-related dates
- Agent responsible for inventorying collection
- Assessment information”

(Digital Preservation Network 2018, Module 2, Slides 17-18)

The assessment step is particularly vital here—rather than simply asking which collections NAL has on various media, the group should determine the long-term value of collections. In preservation efforts, the team should focus on those materials of high value to NAL stakeholders.

Through a combination of technical, workflow, and organizational changes, NAL is poised to provide robust preservation of digital objects. By focusing sustained attention and resources on this important challenge, NAL will be able to confidently meet its mandate to safely steward valuable agricultural information.

Appendix 1 Project Charter

November 16, 2016

Morgan Daniels
Postdoctoral Fellow for Digital Preservation
University of Maryland College Park
mgd@umd.edu

This charter describes a project examining digital preservation workflows across departments at the National Agricultural Library. In accordance with the OneNAL vision for the library's future, the research team at the University of Maryland iSchool (namely Ricky Punzalan, Morgan Daniels, Katie Gucer, and Adam Kriesberg) perceives digital preservation as a concern that transcends departmental distinctions within an organization. One digital preservation infrastructure with accompanying workflows, can, and should, serve the library regardless of the unit responsible for particular content.

In order to help NAL streamline the digital preservation process, the team (lead in this particular effort by Postdoctoral Fellow Morgan Daniels) proposes an assessment of current digital preservation activities across the library. Using a Contextual Inquiry* approach, Daniels will sit down with individual members of each unit in the library to learn about their current digital preservation work, encompassing born-digital, digitized, and web-hosted database collections. By observing individuals as they work in their normal context (at their own workstation) and asking questions as the work proceeds, Daniels will gain an understanding of each person's practices, combining those practices into models that will create an understanding of digital preservation across the library. A number of questions can be addressed through this process, including how does information flow through each department in the process of storing and saving digital materials for the long term? What blockages exist, and how might they be repaired? How can the work be reconfigured to make people's jobs easier? These questions can be answered using Contextual Inquiry methods, which consist of two-hour meetings with individuals from various departments whose work encompasses or interacts with digital preservation activities in some way. Meetings will begin with a brief set of interview questions, followed by an observation period during which Daniels will ask individuals to walk her through their storage and preservation workflows, asking questions about their work along the way. She will ask participants for their permission to audio record the meetings, while assuring that workflows that will be discussed but individuals will not be identified in reports.

By combining the results of numerous such interviews, Daniels will be able to create a big picture view of digital preservation across NAL while retaining the smaller differences between activities in different units of the library. The research will culminate in a report back to NAL, with specific recommendations for redesigning and improving preservation workflows, and in journal publication(s).

The greatest barrier to conducting a Contextual Inquiry project within an organization is getting buy-in across units. In order for this proposed research to succeed, Daniels will need assistance from the heads of each departmental unit in securing the participation of staff members across NAL.

Timeline

Date	Activity
November 2016	Project charter authored, negotiated, and agreed upon by the UMD and NAL teams
December 1-15 2016	Morgan Daniels will develop specific study methods, including semi-structured interview protocol, recruitment text, and contextual inquiry methods
December 15-30 2016	UMD team will update Institutional Review Board (IRB) documents to reflect new study procedures, working with the University of Maryland IRB to assure that the study continues to meet human subjects protections
January 1-March 31 2017	Morgan Daniels will recruit participants for, and perform contextual inquiry interviews with two or more individuals in each NAL unit, creating a workflow diagram for each individual interview. While the individual interview recordings and models will not be made available to NAL leadership (to protect participant privacy), consolidated models illustrating activity within each unit will be created during the next project period
April 1-June 31 2017	Daniels will analyze data during this period, combining individual workflow models into one consolidated model for each unit and an overall model for the entire library
July 2017	Daniels will author a report to NAL describing the study's findings
August 2017	Morgan Daniels' appointment concludes, unless extended

Responsibilities

Morgan Daniels, Postdoctoral Fellow in Digital Preservation at UMD, will design and implement the study, with input from Ricky Punzalan, Adam Kriesberg, and Katie Gucer. She will collect, manage, and analyze study data, derived primarily from Contextual Inquiry interviews. She will be lead author on the report to NAL and on publications and presentations related to this work.

NAL unit leaders will assist the study by participating in Contextual Inquiry sessions, suggesting potential participants, and reaching out to staff members within their unit to request their participation.

About Contextual Inquiry

Contextual Inquiry, a data collection technique associated with user experience research, was developed by Karen Holtzblatt and Hugh Beyer. As a Graduate Student Instructor at the University of Michigan, Daniels taught this technique to Master's students and guided their project teams as they worked to solve an organization's information problem. This project will use techniques described in the book *Rapid Contextual Design: A How-to Guide to Key Techniques for User-Centered Design* by Karen Holtzblatt, Jessamyn Burns Wendell and Shelley Wood. ISBN: 978-0-12-354051-5

Appendix 2 Interview Protocol

The goal of these interviews is to learn about digital storage and preservation related work ongoing at the National Agricultural Library, across all units and departments. Information gained from these interviews will inform recommendations for a digital preservation program at NAL.

Demographic

1. Please state your official title and scope of responsibilities at NAL
 - a. Probe for contractor status, length of service

Extent of digital materials

2. Please describe the types of digital materials you deal with in your work at NAL (including “born digital” materials, digital representations of analog material, and web hosted databases).
 - a. What is the general extent of each type?
 - b. What kinds of growth patterns are you seeing annually?

Digital workflows

3. Please walk me through your workflow for each type of material, showing me the steps you take with each type of material.
 - a. How does the material reach you?
 - b. In what ways do you process it?
 - i. Probe specifically for: reformatting, adding or changing metadata, changing storage location and *specific tools and software used*
 - c. What happens to it after you have processed it?
 - d. Where does it get stored? Is there backup storage/ a redundant copy made?
 - e. What challenges arise during your work receiving, processing, and storing digital materials (if any)?
 - f. What would make this process work better for you? For your colleagues?
4. What metadata standards are most relevant for your work with digital objects?
5. What are the most important aspects of the digital materials you work with to preserve for the long term?

NAL digital context

6. (If applicable) What projects have you previously worked on at NAL related to digital objects?
7. What formal and informal training have you received which prepared you for work with digital materials?
 - a. Probe for grad school, learning from others at NAL, what opportunities would you like to see at NAL, access to training opportunities (finding, time off, etc)
8. What do you see as the challenges on the horizon for NAL and USDA around digital assets? For the agricultural research community more broadly?
 - a. Prompt for digital preservation, safekeeping, backups, archiving, sustainability
9. Who else should I talk to about these issues, specifically people in your unit or people involved in your workflow for these materials?

References

Digital Preservation Workflow Curriculum. Retrieved from <http://dpn.org/members> (in the Best Practices section of the site)

Fedora (2018). Fedora and Digital Preservation. Retrieved from <http://fedorarepository.org/fedora-and-digital-preservation>

Holtzblatt, K., Wendell, J. B., and Wood, S. (2005). Rapid Contextual Design: A How-to Guide to Key Techniques for User-Centered Design. Morgan Kaufmann Publishers, San Francisco, CA.

Kriesberg, A. (2016). NAL Digital Curation Plan. Beltsville, Maryland. (Internal report, not published.)

Life Cycle Assessment Commons. (n.d.) Life Cycle Assessment. Retrieved from <https://data.nal.usda.gov/life-cycle-assessment>

Phillips, M., Bailey, J., Goethals, A., Owens, T. (2013). The NDSA Levels of Digital Preservation: Explanation and Uses. *Proceedings of the Archiving (IS&T) Conference*, April 2013, Washington, DC. Retrieved from: http://ndsa.org/documents/NDSA_Levels_Archiving_2013.pdf

Poelchau, Monica, et al. "The i5k Workspace@ NAL—enabling genomic data access, visualization and curation of arthropod genomes." *Nucleic acids research* 43.D1 (2015): D714-D719.

Rieger, T. ed., (2016) Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Files. Federal Agencies Digital Guidelines Initiative (FADGI) Still Image Working Group. September 2016. Retrieved from http://digitizationguidelines.gov/guidelines/FADGI%20Federal%20%20Agencies%20Digital%20Guidelines%20Initiative-2016%20Final_rev1.pdf