Migration Misadventures
The Challenges of Creating an Inter-Institutional Migration Plan

Laura Kathryn Nicole Jones

School of Information Studies, University of Wisconsin – Milwaukee[1]

Abstract

Archival institution migration projects are planned data transfers from existing data formats to more accessible formats that assist in preserving transferred data, improving information retrieval, and preventing obsolescence. Various archival institutions are increasingly planning migration projects because current technology makes the process easier and more affordable than in the past. In this proceeding, I will walk through the planning stages and associated challenges of NASA's inter-institutional migration plan, Archival Description Center Migration Plans Project, which is part of a larger reorganization of the archives' functional area to an Enterprise program. For this project, we utilized the open-source, standards-based, web-based applications Access to Memory, Archivematica, and OpenRefine. At the current project stage, we have collaborated with archivists from seven of the eleven existing NASA archive centers. One of the goals of this Enterprise reorganization is to facilitate cross-collection search and discovery and establish cross-institution metadata standards.

Introduction

According to Dingwall (2017), archival migration "is the conversion of a file from a format at risk of becoming obsolete to a more current format" (p. 140). The original intent of migration projects was to prevent information loss due to obsolescence, described by Dingwall (2017) as "the [in]ability to access the content of files of a particular format" (p. 139). However, the advantages of providing information about an institution's holdings online have added new benefits to migration plans: researcher and patron ease of access, improved information retrieval, cross-collection search and discovery, and technological interoperability.

NASA Archives is currently undergoing restructuring into an Enterprise archive program. One goal of this restructuring is to standardize and connect all Center archives through the Archival Description Center Migration Plans Project (Migration Plans Project). Before the Migration Plans Project, each Center archives organized and described its materials uniquely, making inter-institutional information retrieval difficult.

The Migration Plans Project aims to design, test, and develop standardized workflow directions that all Center (NASA locations with archives) archives will eventually use to process, describe, and import metadata about holdings and digital objects into one central data management system. The system we are using is Access to Memory (AtoM) (see section Materials and Methods: Software). Once metadata and digital objects are imported into AtoM, they will be accessible online for inter-institutional use and, classification pending, available for viewing online by the public.

There are currently eleven Centers with archival collections: Ames Research Center, Armstrong Flight Research Center, Glenn Research Center, Goddard Space Flight Center, Jet Propulsion Laboratory, Johnson Space Center, Kennedy Space Center, Langley Research Center, Marshall

[1] I would like to acknowledge the Wisconsin Space Grant Consortium (WSGC) for financial support and thank them for providing me with great opportunities.

Space Flight Center, NASA Headquarters, and Stennis Space Center (Fig. 1). Of these eleven, we have worked with seven at our current stage in the Migration Plans Project: Ames Research Center, Glenn Research Center, Goddard Space Flight Center, Jet Propulsion Laboratory, Marshall Space Flight Center, NASA Headquarters, and Stennis Space Center.

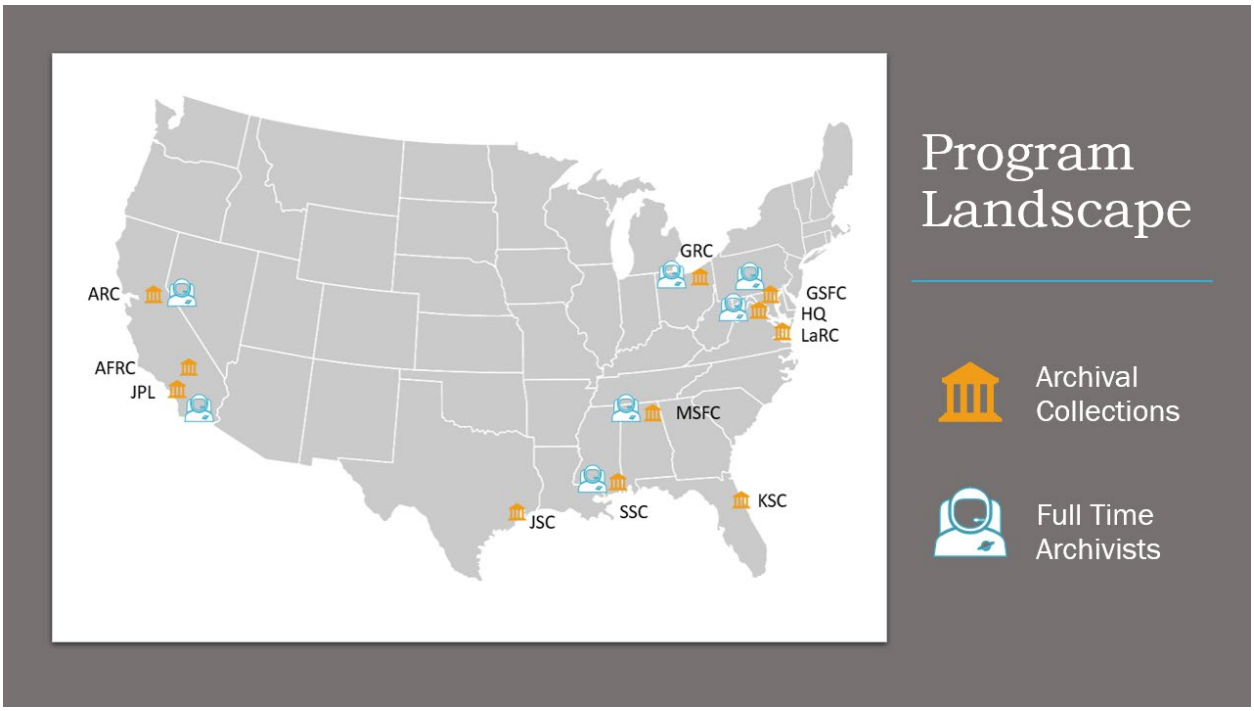

Figure 1: Map of all NASA Center archives (McIntyre, 2021).

Materials and Methods

Due to the breadth of the Migration Plans Project, not all Center archives are currently involved with the Project. We began with those who already begun individual organization projects to create standardization across their own collections. These included Ames Research Center, Glenn Research Center, Jet Propulsion Laboratory, Marshall Space Flight Center, NASA Headquarters, and Stennis Space Center. Goddard Space Flight Center was the guinea pig for the Migration Plans Project because they already had an AtoM webpage and importation workflow (see section Step Three: Write Workflows).

Software. The three software programs we are using in the Migration Plans Project are AtoM, Archivematica, and OpenRefine. AtoM is specifically designed to help archival institutions manage, describe, and access their holdings across repositories through the Web. Archivematica works in tandem with AtoM to ensure that digital materials are preserved in ways that will prevent them from succumbing to obsolescence. After materials are run through the Archivematica program, they are then linked to their descriptions in AtoM to be accessed by internal or external users. Both programs are open-source, standards-based, web-based applications created by the company Artefactual. The open-source and multi-repository nature of these programs made them high contenders for use by NASA's History and Archive Branch. There are drawbacks to open-source software, however.

Because open-source software is free, it is usually unpolished with a high learning curve. Users typically must build a lot of their own content and conduct extensive testing and research to mold the software into the desired format. The positive of this, however, is that the software is malleable enough to fit nearly any institution's desired specifications and features. This flexibility will help us create a system tailored to the needs of NASA's large multi-repository nature and varied holdings.

The third program we are using is OpenRefine. Similarly, to AtoM and Archivematica, OpenRefine is an open-source, web-based application; however, OpenRefine was developed by Freebase and then Google before becoming an open-source program. OpenRefine is specifically used for standardizing large amounts of data. The program makes finding multiple renditions of terms quick and allows for changing all incorrect instances to the proper authority terms (a specific word used to describe archival records to standardize language for improved information retrieval) simultaneously. In the workflow, this program is used to standardize all data spreadsheets before they are imported into AtoM or Archivematica (see section Step Three: Write Workflows).

Step One: Obtain Holdings Information. The first step in the Migration Plans Project was to determine the amount of each format type held at Ames Research Center, Glenn Research Center, Jet Propulsion Laboratory, Marshall Space Flight Center, NASA Headquarters, and Stennis Space Center. We accomplished this step through several sets of interview questions.

My co-intern, Robin Klemm, and I both completed this step of the Migration Plans Project. We were each given different Centers to contact. Klemm emailed Ames Research Center and Jet Propulsion Laboratory; I emailed Glenn Research Center, Marshall Space Flight Center, NASA Headquarters, and Stennis Space Center.

Our original emails were requests to contact the archivists at their respective Centers about their holdings and invitations to interview each archivist in their preferred method. My contacts all chose to communicate through email. The second round of emails contained broad questions about each Center's holdings. We thought these broad questions were clear enough to indicate that we wanted specific information about each type of format and authority, but the archivists responded with generalizations about their holdings.

The information we had obtained thus far was not specific enough for us to determine the types of workflows necessary for the Migration Plans Project, so the lead, acting chief archivist Holly McIntyre, drafted a set of more specific questions that I adapted for different formats we knew were available at each Center (see Appendix I). This questionnaire, with the support of follow-up questions or clarifications, provided more complete holdings information that allowed us to move forward in the Migration Plans Project.

Unfortunately, many of our contacts were very busy, and we had to wait multiple days or weeks for responses to each email. This delayed the timeline of the Migration Plans Project significantly. My suggestion to future archivists, and questionnaire writers in general, is to

begin the interview process with the most specific and detailed questions you can devise before asking more general questions. We also discovered that there was some confusion over the nature and goals of the Migration Plans Project. If a more formal presentation on the objectives of our questions and how they aligned with the Migration Plans Project had been presented to all participating Centers, perhaps fewer clarifications or follow-up questions would have been necessary.

Step Two: AtoM Testing. Information is imported into AtoM through a comma-separated value (CSV) file, typically through the user interface. However, due to the open-source nature of the software, some information can only be imported through the command line. To test these particulars, I was assigned the side project AtoM Container List Workflow, while Klemm and I were contacting our respective archivists.

The goal of the AtoM Container List Workflow project was to determine the easiest way for archivists to upload physical location information into AtoM. At this time, Goddard Space Flight Center was not including physical location information in their CSV imports and needed a way to add this missing information.

With assistance from our other supervisor, Goddard Space Flight Center archivist Christine Stevens, I researched and tested different methods of importing physical location information. This process involved reading AtoM's documentation, browsing and posting in the community forum, leading a test run with the Goddard Space Flight Center's library systems administrator, and testing different ways to input information in imported CSVs. We concluded that the physical location information should be included in initial CSV imports but can also be added to existing descriptions in AtoM via the command line using *roundtrip*. The *roundtrip* command edits an existing CSV in AtoM by exporting the specific file, editing it, and then re-importing it.

Step Three: Write Workflows. This stage of the Migration Plans Project is writing the actual workflow documentation. Klemm and I began this process by breaking down the formats we discovered in Step One. We created individual outlines for each using the Microsoft Teams Wiki tab, then drafted the process each format must go through to become standardized and then transformed into a compatible CSV file for AtoM importation.

The basic steps of the workflow are as follows:

1. Obtain or create a spreadsheet of holdings data at each level (e.g. collections, series, subseries, folder, item).
2. Standardize each spreadsheet according to the authority file (reference document that provides all authorized versions of terms).
3. Reference workflow documentation by media type to determine exact importation steps since different media types have different requirements within the Archivematica program.
4. Import data into AtoM.

Once the format outlines were completed, we created a living workflow document on the Archives' Confluence page using the Goddard Space Flight Center's AtoM importation

workflow as a base. As more archives begin their standardization process and bring other formats to the table, those formats will be integrated into the workflow.

In addition to the workflow, Klemm and I wrote individualized outlines based on McIntyre's questionnaire, detailing the specifics of each Center's holdings. Once the basic workflow was written in Confluence, these Center outlines were reworked to provide individualized summaries for each Center as well as curated migration plans based on material prioritizations (discussed in the next section).

Step Four: Prioritize Materials. After the Center outlines were completed, Klemm and I used these to determine migration priorities for each Center. Part of this process required contacting the archivists we interviewed in Step One to obtain their input on the importance and popularity of collections. The three criteria we used to evaluate collections were: 1) volume, 2) usage, and 3) value.

Step Five: Archivematica Testing. The final step in the Migration Plans Project was editing the workflow to include instructions for importing digital objects into Archivematica and then linking them to their respective AtoM descriptions. This required researching Archivematica documentation and then testing methods to correctly document the entire process step-by-step. Those steps were added to the completed workflow in Confluence.

The advantage to creating the workflow as a living document is that we could simultaneously work on multiple steps, and we will be able to edit the document as methods change in the future, rather than beginning the writing process over for each new iteration. Because the workflow will be stored in Confluence, it will also be interactive and easily accessible to all current and future archivists.

Discussion

Based on my experience with the Migration Plans Project, I have three suggestions for archivists attempting a similar project: 1) write detailed and specific interview questions, 2) create your workflow as a living document in a program like Confluence or Notion, and 3) plan extra time into your project timeline. The biggest mistake we made with the Migration Plans Project was not budgeting enough time for mistakes and other peoples' time. When planning a project that involves people who are not directly assisting with the project, remember to plan extra time for their input or responses; many professionals are very busy, and it is very kind of them to assist with projects external to their own responsibilities. Respect their time by planning ahead and budgeting time for their responses.

Even though we had setbacks with the Migration Plans Project, overall, results and progress have been positive. The delays turned out to be lessons in disguise that helped us improve the Migration Plans Project and our methods of implementation.

The Migration Plans Project was initially slated to end with the completion of the workflow on August 12, 2022; I am thankful that Klemm and I were given the opportunity to continue this work through to its completion on December 05, 2022. We learned several new skills from this

experience and improved on existing skills. This was our first exposure to AtoM, Archivematica, Confluence, migration planning, and workflow creation; our experiences in this internship expanded our skills in collaboration, writing, controlled vocabulary, and metadata. Even though we had a few misadventures on our journey, it was more of a positive learning experience than a negative one. The next steps for the project team are 1) to bring on new interns, 2) train interns and NASA archivists on AtoM, Archivematica, and OpenRefine, and 3) train users on the workflows so the standardization and importation of holdings data can begin. The entire process will likely take several years, but this is not unusual for archives of such magnitude. The sun is setting on this quest and will soon rise on a new (mis)adventure.

Acknowledgements
I would like to thank the Wisconsin Space Grant Consortium for funding my internship twice; the National Space Grant College and Fellowship Program for both internship opportunities; Holly McIntyre and Christine Stevens for their mentorship and guidance; Robin Klemm, my fellow intern; Rachel Owens and Megan Goller for helping me navigate paperwork; Bob Arrighi, Sarah Jenkins, Jordan Whetstone, and Jessica Herr for providing information about their NASA Center holdings; the University of Wisconsin – Milwaukee for my education; and Kelli Bogan, my Society of American Archivists mentor. Thank you.

References
Dingwall, G. (2017). Digital Preservation: From Possible to Practical. In H. MacNeil & T. Eastwood (Eds.), *Currents of Archival Thinking* (2nd ed.) (pp. 135-161). Libraries Unlimited: An Imprint of ABC-CLIO, LLC.
McIntyre, Holly. (2021).

Appendix I
Example Follow-Up Questions for Enterprise Migration Plan
Spreadsheet Inventories:

1. Are the spreadsheet inventories electronic?
2. Where do they exist?
3. How many are there?
4. How many rows are in each?
5. Which descriptive standards and controlled vocabulary/authority records are employed across them?
6. Please tell me the number that use each descriptive standard/vocab/authority.
7. Is there any supplemental information to better understand the way that this vocabulary/authority was used? Please include inconsistencies and loose adherence.
8. How does this vocabulary/authority exist?

Electronic Finding Aids:

1. Where do they exist?
2. How many are there?
3. Which descriptive standards and controlled vocabulary/authority records are employed across them?
4. Please tell me the number that use each descriptive standard/vocab/authority.

5. Is there any supplemental information to better understand the way that this vocabulary/authority was used? Please include inconsistencies and loose adherence.
6. How does this vocabulary/authority exist?

JPG Access Copies

1. Where do they exist?
2. How many are there?
3. Do they have corresponding preservation formats?
4. Are they born digital or digitized?
5. If digitized, do the originals still exist?
6. What description formats/standards are associated with them?
7. How many access copies are associated with each format/standard?
8. Is there any supplemental information to better understand the way that this vocabulary was used? Please include inconsistencies and loose adherence.
9. How does this vocabulary/authority exist?
10. Are there any file naming conventions used?
11. Please include examples and quantity, include files that do not follow naming conventions (e.g. EpsonIMG003.jpg).
12. Is there any embedded metadata?
13. Please include examples and quantity, include files that do not have embedded metadata.

PDF Access Copies

1. Where do they exist?
2. How many are there?
3. Do they have corresponding preservation formats?
4. Are they born digital or digitized?
5. If digitized, do the originals still exist?
6. What description formats/standards are associated with them?
7. How many access copies are associated with each format/standard?
8. Is there any supplemental information to better understand the way that this vocabulary was used? Please include inconsistencies and loose adherence.
9. How does this vocabulary/authority exist?
10. Are there any file naming conventions used?
11. Please include examples and quantity, include files that do not follow naming conventions (e.g. EpsonIMG003.jpg).
12. Is there any embedded metadata?
13. Please include examples and quantity, include files that do not have embedded metadata.

Misc.

1. Are there any description or access copy formats that were not included in the original response to my questions? Please consider legacy and analog collections and anything else that may or may not be a current work in progress.
2. Please include the same information from the questions above for each individual description format and access copy format.