5-18-2018

An Overview of Planning and Implementing Large-Scale Digitization

Emily Lapworth
University of Nevada, Las Vegas, emily.lapworth@unlv.edu

An overview of planning and implementing large-scale digitization

Created March 8, 2018 by Emily Lapworth for the Nevada Statewide Large-Scale Digitization Workshop

Table 2. Framework for a Digitization Project

Creating digital collections encompasses a diverse array of activities. The list of main functional areas that follows represents a wide range of skills:

Selection
- material selection based on research, learning, and teaching needs
- copyright-status assessment

Requirements analysis to set technical requirements for:
- digitization
- metadata
- access and use
- other repurposing areas (e.g., print on demand)

Preparation
- conservation, disbinding, tagging
- physical volume organization by content or format type

Digitization
- digitization (in-house or outsourced)
- image processing
- creation of archival and derivative files
- structuring

Quality control
- development of a QC strategy
- selection of QC tools
- development of assessment workflow
- plan for correcting and reintegrating unacceptable images and other deliverables

Metadata
- descriptive, structural, administrative, preservation
- controlled vocabulary, taxonomies, ontologies
- selecting and implementing standards for interoperability, discovery, etc.
- file-naming conventions and persistent IDs
- OCR

Technical development
- repository and storage plan
- digital content delivery platform (image database)
- discovery and navigation tools
- Web services
- Web design and development

Project management
- workflow coordination
- financial management
- assessment and usability analysis
- promotion
- user support

Life cycle management
- preservation strategies and procedures
- ongoing content, metadata, application revisions, additions, etc.

Rieger, Oya Y. "Preservation in the Age of Large-Scale Digitization: A White Paper." *Council on Library and Information Resources*. February 2008, p. 16. https://www.clir.org/pubs/reports/pub141/pub141.pdf.

1. Identify collections for digitization
    a. Brainstorm your goals for this project. Think about what you will do with these digital surrogates, and who your audience is.
    b. Criteria for selection of materials
        - Formats: Start simple. If everything's the same, large-scale workflows are easier to apply. Ultimately you will need to create different workflows for each format with differing requirements. For example, print photos are digitized differently than film negatives. Text documents benefit from OCR, while photos do not, and handwritten materials present additional discoverability challenges. When creating complex digital objects with different formats within them things can become even more complicated.

        - Condition: fragile materials require extra handling time and possibly additional physical treatment prior to digitization
        - Existing arrangement and description: it is easiest if online access can directly mirror physical access, but the materials may need additional arrangement and description before digitization, depending on your goals. If the materials already have item or folder level description that is ideal. If there is any hierarchy in the existing description, especially inconsistent or complex hierarchy, consider how you will reuse that description for digital objects.
        - Copyright: plan on providing public online access only if you own the copyright, have permission from the copyright holder, or if it is a strong case of fair use.
    c. See preparation step (below) to come up with some idea of how you will undertake this project. It will likely be modified during the actual preparation, but you need to have some idea of what you will do and how you will do it in order to gather support and resources.
2. Assess the technical infrastructure needed to create, manage, provide access to, and preserve the digital files.
    a. Estimate how much storage space you will need, and how much space will be needed for long-term digital preservation.
    b. Make sure that your current digital preservation policies and workflows will be able to accommodate this project. Adjust them if needed.
    c. Identify what equipment and software will be needed and if you already have it, can acquire it, or can use someone else's.
    d. Assess if your existing workflows and systems for providing access to digital materials will be able to accommodate this project, and what changes you might need to make.
    e. Technology could be a great area for collaboration! If you lack certain resources, explore opportunities to collaborate with other institutions.
3. Coordinate with other stakeholders to verify choices and plans for digitization.
    a. Find out what kind of support there is (financial, staffing, etc.) from management, administration, and the community.
    b. Identify possible collaborators and discuss plans, make agreements, etc.
    c. Decide who will manage and oversee the project and how different responsibilities will be distributed.
    d. Identify and apply for grants if appropriate.
4. Prepare collections for digitization
    a. Arrangement: assess how are the materials physically arranged and described, and if it will help or slow down your anticipated workflows. Plan for and complete additional processing if needed.
    b. Decide how you will display digitized materials. Mirroring existing arrangement is the easiest, but you also have to consider the file formats you want to create.

    c. Description: figure out how you can reuse existing description. Plan metadata fields, vocabularies, prioritized subject terms and names.
    d. Prepare preliminary metadata. Reuse what you already have!
    e. Prepare physical materials. Verify that physical contents of the collection match existing description or inventory. Remove staples, unbind, unsleeve, flatten, etc. Identify and address any preservation or conservation issues.
    f. Identify physical formats *(this will help determine timeline and what equipment is needed)*
    g. Decide: outsource or in house?
    h. Create and test workflows and procedures
    i. Create documentation for workflows and procedures *(important for duration of project, for reusing for future projects, and also for future employees stewarding these digital assets to know what you did and how you did it)*
    j. Create and prepare systems, documents, or mechanisms to track work *(important to stay organized, especially when dealing with a large amount of materials or a team of workers)*
5. Digitize collections
    a. Set up consistent file naming procedures and make sure they are followed.
    b. When dealing with mixed materials in house: Depending on equipment and composition of materials, start with the easiest or what you have the most of, then take note of other formats (e.g. transparencies, oversize, etc.) that require different equipment or settings so you group them together to do later all at once.
    c. Keep specifications simple if possible, especially if you have student workers doing the digitization. *(For example, if you have complex digital objects with both text and photographic prints, and can digitize both materials on the same equipment without changing settings, do so. If you normally digitize text at 300 ppi but want photos at 600 ppi, rather than having the technician stop and change the settings, capture all at 600 ppi if you have the space.)*
    d. Auto-crop is a great tool if you have it but otherwise try to improve the efficiency of your processes with any tools at your disposal. Sometimes this can be as simple as placing the item with the correct orientation to avoid the need to manually rotate later.
    e. File formats: Archival images are generally tiffs. Smaller derivative files may be necessary for access or to speed up OCR processes. Sometimes it's better to output them at the time of scanning than to batch process later.
6. Image processing
    a. See above: try to improve your digitization workflows and procedures to shave time off of image processing.
    b. OCR: If you have textual materials OCR makes them much more accessible with less manual work into creating detailed metadata. This is especially true for large aggregations of textual documents. Resist the urge to have perfect OCR. Something is better than nothing, and when dealing with scale, you do not have

the time to correct everything. Here is also an opportunity for crowdsourcing, if you have the technological resources to set it up.

c. OCR file output: depending on how you choose to display and make the digital surrogates available, you may need to output text files and/or PDF/As.

7. Description and access
    a. Reuse description that already exists (e.g. from an inventory or a finding aid). If a finding aid exists, make sure you are using all available information and understand how description is inherited and can be reused.
    b. At the beginning of the project transform the metadata that already exists into a format you can use to describe the digital objects. You can add to this existing metadata throughout the workflow.
    c. At the beginning of the project identify preferred subject terms and important names to look out for and add to digital object metadata when appropriate. This is especially important when metadata is created by students or teams or anyone unfamiliar with the subject matter of the collection. It will help ensure consistency and make faceting better for users.
    d. Explore how search engine optimization (SEO) works for your public online access system. Take that into consideration when creating metadata in order to optimize discovery of the materials.
    e. Make it as easy as possible for users to identify the provenance of the digital object and to find other digital objects from the same collection.
    f. Consider the links between the original collection description and the digital surrogates. Consider adding digitization information or links to digital surrogates into finding aids and other records. Consider also adding a link to the finding aid into the digital object metadata. Consider using persistent identifiers, such as ARKs (Archival Resource Key), to do this, instead of using regular URLs.
    g. Find out how your access system indexes full text transcripts and how it displays different file formats. Consider if you are you able/ if you want to offer multiple file formats of a digital object. For example, a compound digital object that includes both text and images could be available as a collection of image files, a single PDF file, or both. Identify what would be most useful to your users.
    h. Don’t forget about structural, administrative, technical, and preservation metadata!
8. Quality control
    a. Have a strategy (e.g. sampling), guidelines, and goals for QC.
    b. For staff performing quality control, identify the most important things to look for.
    c. Decide how much time should be spent on QC.
    d. Identify and acquire any automated tools that can be used.
    e. Set up procedures or steps to follow when errors are found.
9. Digital preservation
    a. You should have already planned how you will ensure access to and preservation of the digital files and metadata in the long term. Best practice is to have policies in place identifying what digital assets should be preserved and to

what extent. Identify applicable standards and best practices, implement software and technical solutions.

b. Set up workflows and procedures to ensure that the digital files receive appropriate ongoing digital preservation treatment.

10. Publicize and promote
    a. Work with administration, collaborators, and other stakeholders to publicize and promote the project.
    b. Depending on your audience, social media, academic listservs, and professional organization publications can be other avenues to spread the word.
    c. Set up harvesting with Mountain West Digital Library for inclusion in the Digital Public Library of America.
11. Assess
    a. Web statistics can be used to track the use of online materials. See SAA/ACRL's Standardized Statistical Measures and Metrics for Public Services section 8 "Online Interactions" for general information on what information to collect, and the Digital Library Federation's Best Practices for Google Analytics for specific information on Google Analytics.If you are a CONTENTdm user, see Getting Started with Google Analytics in CONTENTdm.
    b. The Toolkit for the Impact of Digitised Scholarly Resources is another resource for learning how to measure the impact of digitization projects. It includes information on methods beyond web analytics, such as surveys and altmetrics.
    c. Record and compile any oral or written feedback received from stakeholders and audiences.
    d. Analyze feedback and use statistics to identify areas of success and areas for improvement. Make improvements as necessary and incorporate findings into planning for future projects.