The workflow of the digitization process for archival collections first became an interest of mine as I worked for Rachel at the University of Louisville in the Photographic Archives, and currently, with my work in the digital archives at the Wisconsin Historical Society. My work at both institutions has involved taking large collections and helping move them through the digitization process. My focus today is talking about metadata, particularly when metadata should be created in the digitization process so that it is done most efficiently, and then how this can be done to also improve the accessibility of these collections once online.

Engaging in More Product Less Process (or MPLP) is an imperative that many institutions face because of limited amount of person hours they have to process collections and make them available. MPLP is a mantra at the Historical Society because the sheer volume of their accessions and their mission to put collections in the hands of interested patrons as immediately as possible far exceeds their available labor, even with dozens of LTE and student employees.

MPLP is the opposite of detailed, item-level processing because it foregoes the item-level work on the collection in favor of doing basic processing on a greater number of collections. So, applying MPLP to the process of digitizing an archival collection might seem, at first, to be absurd because MPLP is about streamlining processing. The very act of digitization demands intensive item-level work.

If we think of MPLP in terms of mass digitization, however, and apply this understanding to the digitization project workflow, we can see where significant efficiencies exist in the mass digitization process that can allow for streamlined processing. But first, I wanted to briefly reclaim the term mass digitization. It has become synonymous within the academic librarian field with mega-scale projects like the Google Books mass digitization effort.

Librarians like Karen Coyle in her article Mass Digitization of Books, remind us, however, that “Mass digitization is more than just a large-scale project. It is the conversion of materials on an industrial scale.” In other words, the defining feature of mass digitization is its use of an “assembly-line” process and other labor efficiencies to produce more product, with less effort. Mass digitization has the same underlying sense of a streamlined process as MPLP; it is intended to be quick, efficient, and provide the means to produce a viable product (here the
readying of a collection) with the minimal amount of labor in the quickest amount of time.

Because item level work is absolutely unavoidable in the digitization of a collection, it’s important to take advantage of the time spent doing item level work by implementing a process that uses this time as effectively as possible. This is true whether the work being done is a page count of the entire collection to assess and map out a project timeline, or the scanning of each individual item into digital format.

One of the greatest efficiencies can be realized by performing subject tagging simultaneously while counting or scanning the documents. Having the same workers who did the page count or scanning also input their subject tags, notes and other metadata that has been collected doing item-level work into a spreadsheet or database.

This accomplishes several things.

- First, it cuts down on the amount of time and handling of a collection that must occur.
- Second, concurrent scanning and tagging eliminates the need for the item (which may be fragile) to be accessed multiple times.
- Third, it eliminates the extra time (and errors that occur) when an item must be repeatedly handled by the same (or different) students or employees who must re-familiarize themselves with an object when scanning and tagging occurs days, weeks, or months apart.
- Lastly, inputting metadata concurrently with item level work helps cut down on the number of re-interpretations that must occur with an item and its metadata. In other words, workers don’t have to take dormant data that might have been created months ago by a worker who has already moved on, and rely on it to continue the item-level processing. The item and its metadata are complete and ready for the next step in processing.

Obviously, a significant barrier to this method is ensuring that people working on the digitization project understand the metadata in the same fashion. This does not always happen on its own accord and requires a bit of calibration, but the
benefits of the extra time spent in meetings still outweighs the costs because this time spent reduces:

- The time each worker needs to spend trying to interpret metadata from a scanned image.
- And, the time spent correcting errors when ambiguous terms are used different ways by different workers or workers re-approaching dormant metadata.

------------------------------------------

Reducing errors and creating efficiencies by inputting metadata during the steps of item-level processing is one way that the spirit of MPLP can be used in the mass digitization of digital collections. Creating unique subject headings is another effective way of making underrepresented collections, or any collection for that matter, more accessible to patrons.

Basic administrative headings like personal names and geographic names continue to utilize Library of Congress subject headings because they are more stable and likely to be consistently used the same way across collections and institutions.

Subject headings are more complicated because they require a large degree of subjectivity on the part of the cataloger and sometimes use awkward or outdated language that doesn’t match the search terms end users would be likely to employ. In digital environments, they improve patron’s capabilities of finding what they need when they may not have ready access to an archivist to consult with on their searches.

At the extreme is the social tagging or “folksonomy” approach, which allows users to provide their own search terms. This is a solution plagued with problems, because users have to find the materials in the first place in order to tag them, and because using a completely uncontrolled vocabulary does not enable cross-searching. Furthermore, misspellings and personal agendas will certainly be issues here.
WHS has employed a middle-ground solution, which we call “Grassroots” or “bottom up” subject headings. These are local, controlled terms that use more colloquial language than LCSH but have more consistency than social tags. The relatively short array of options makes it more efficient to apply these terms, whether at the item, folder, or collection-level.

“Bottom up” subject headings use a customer-focused approach that generates keywords by consulting the indexes of books used by patrons as primary sources. Bottom up tagging also uses keywords from other collections that have already been digitized within the institution and integrates these collections’ subject headings into present and future projects. Using keywords that have already been created for specific topics helps keep the digitized collections connected and offers better search and retrieval results.

I think it’s important to recognize that “Grassroots” headings are not a cure all. Using them has significant flaws, like local idiosyncrasies in the use of terms (“Southern Mentality”). “Grassroots” subject headings also never truly having the ability to fully anticipate what users want, there will always be gaps.

As Sue will discuss next, the level and extent to which any controlled vocabulary terms are applied is a decision you’ll have to make about efficiency versus accessibility, and if providing access to materials featuring underrepresented groups is a goal, there are tradeoffs in efficiency.

For example, while at the University of Louisville I cataloged a large collection of portraits from the early 20th century. I used city directories, the Encyclopedia of Louisville, and local “Who’s Who” volumes to identify the first names and, often, maiden names of many of the women pictured, rather than identifying them merely as “Mrs. So-and-So” as they had been labeled by the donor.

Rachel and the reference staff in Special Collections can attest that people are finding these photos because they are contacting UofL to order copies, provide additional information, and make corrections. Research/genealogist usage of that set of portraits would have been limited to those of the men only had we not taken extra steps to identify the historically underrepresented individuals.