Tool summary, Illinois Wesleyan University, May 2014

Executive summary: Duke Data Accessioner will be used to process digital objects being sent on removable media in the archives. Steps needed to implement Archivematica locally will be explored further. DuraCloud will be recommended for content types determined to be worth the expense (estimated expense currently, $1,500/year for the first TB) of bit-level preservation.

My assessment of each tool/product/service is below. The list includes Archivematica, Bagger, CINCH, Duke Data Accessioner, DuraCloud, MetaArchive, and Preservica. Additional tools reviewed but not fully tested are Curator's Workbench and Hoppla.

Tools tested:

Archivematica (https://www.archivematica.org/wiki/Main_Page) was the most promising and comprehensive of the macroprocessing tools I used. It seemed to do all of the steps recommended by OAIS Reference Model, including file normalization, in a way that could be accomplished by work-study students (my only consistently reliable form of labor). We didn't actually test it all the way through to the DIP stage, but I wish we could have. We stopped at AIP creation which seemed successful, but I didn't have a place to transfer AIP to, so I can't say what would have happened afterwards. A few unresolved questions remain about implications for actions at some of the decision making steps. I wasn't always certain why I was choosing one option or what would happen if I chose another. Maybe embedded references to appropriate user community threads would help? Also, the prompts for making another decision were only visual, so close attention to the processes taking place over long, uneven periods of time would be needed. Finally, I am not certain my institution's IT staff could help me install a virtual box. I don't think I could have accomplished that setup without help from our project IT staff. Others in the POWRR group say there's an update (v1.0) available but that it requires some expertise to get it installed. We've also been told a fee-based hosted service is coming in the future. My criteria for pursuing that version will include cost and transparency of backend storage suppliers.

Update in April 2014: A new version of Archivematica may become available if the developer identifies institutions willing to invest. Aspects listed above that I expect to make it difficult for our institution to use are not on the enhancements' list.

Bagger was easy to use once installed and didn't require any individual, item-level metadata decisions. MetaArchive requires processing with the previous iteration of this tool (Bag-It) but our project (DigitalPOWRR) members tested this newly developed version; we understand it will be widely available soon. The tool "bags" a whole collection for storage or transfer at one time and creates a checksum on this entire package as a whole rather than on individual digital objects a collection may contain. The tool does not require pre-processing steps (like standardizing file

names), but it is important to know if such a step will be required by a storage system the bag is transferred (this was the case with MetaArchive). The download was straightforward, but there were a few set up steps that were not intuitive. I feel confident I could keep using Bagger if needed, but a major improvement would be to have the required metadata fields loaded and ready to use. It would also be helpful if constant data (like institutional name, record creator, etc.) could be stored within the template of the tool. We were given instructions by Educopia ("1.3.1 Creating a Custom Metadata Profile") on how to create such a file but those were not intuitive. A library staff member and I worked on these instructions together and could not complete the steps.

"***CINCH*** (Capture INgest CHecksum) is a tool that automates the transfer of online content to a repository, using ingest technologies appropriate for digital preservation" (http://cinch.nclive.org/Cinch/). This is the easiest Web-archiving tool I've used. It is designed to capture individual files and works with several text and media formats. The down side to using it is that specific URLs must be provided, so it would be labor intensive to compile initially. The greater problem, though, is that our campus content creators often use different file names for content that's merely updated online, so the labor of identifying and entering new URLs would have to be constantly repeated.

Duke Data Accessioner (v0.4.0) is an open source tool designed to move content from removable media (such as disks and flash drives) and place it on a server for further processing (http://library.duke.edu/rubenstein/uarchives/about/data-accessioner). I was able to download and launch this product without difficulty. I intend to use it to process all removable media types until a better workflow becomes possible. DDA creates a manifest that holds file integrity (checksum) information, compares content against known PREMIS standards such as DROID (Digital Record Object Identification) and JHOVE ((JSTOR/Harvard Object Validation Environment). The tool also recreates existing folder structures without changing the transferred objects; therefore, a "master" copy can be held as a migrate-able server file for preservation purposes and a derivative can be created for access, if needed. The manifest is in XML and is stored with the highest level (file or folder) it relates to. XML can be ingested into Excel for ease of analysis by manual sorting.

DuraCloud (http://duracloud.org) was an easy, intuitive product once all the set up was complete. The company manages digital preservation activities like integrity checks and metadata management and outsources backend storage choices. They offer the most affordable, managed options for IWU at this time. The product is open source and run by DuraSpace—their product allows for a lot of metadata but does not require it. Access restrictions can be managed for any part of the ingested collections. They do not normalize or migrate content, but their

features are adequate for our needs and they have an active development community. This was the only preservation company that had data on the number of degraded files that their system has detected and replaced. That number is zero.

There are some parts of the documentation that need to be clarified, but my biggest problem with the product was that the only way to upload multiple files at a time is through the Sync tool, and then every time the computer starts afterwards it checks for new content. I can think of several reasons why this is a good aspect of the tool (one listed below), but it would be annoying for people using one-time bulk uploads for content they don't intend to update (specifically thinking of my previously digitized periodicals collections that won't be added to). The only other issue I had with it was that I operate in a folder-aggregated environment but individual files are the default view on the administrative side of DuraCloud. Maybe something else is possible but I didn't discover it.

I think a big benefit to using DuraCloud would be for streaming media. In our IR, we've made a decision to host metadata that makes media discoverable but store large files on a streaming server to ease user access. DuraCloud makes it possible to store/audit media files and also allows for public streaming access. This is an appealing feature, although the only storage location they offer for this is Amazon. DuraCloud makes three storage backends available to subscribers (two are Amazon products: S3 and Glacier, and one is from the San Diego Supercomputing Center).

I mentioned finding a positive side of the Sync tool…I transferred about 8GB of content from our IR (Digital Commons, hosted by bepress) and observed a benefit of Sync: the process analyzes for duplicate content before uploading, so we could potentially save storage space by not transferring anything that's already there. As noted elsewhere, my IR is not able to identify newly added content and so our only available option is to transfer in bulk on a quarterly basis.

MetaArchive (http://metaarchive.org) is a cooperative run by the Educopia Institute and its members based on LOCKSS preservation principles. The POWRR project used NIU as a node for the partner institutions to test file transfer protocols. We proved NIU could sustain these processes (with one remaining technical question described below) but other questions remain over how fees and legal agreements would be handled. MetaArchive requires that a "find-bad-files.py" script be run against folders before using Bagger and MetaArchive requires the use of Bagger before transferring material into the participant's node (*i.e.*, a server running LOCKSS software). Resolving problems with file naming conventions proved to be a bigger problem for me than anything else in this project. A POWRR colleague shared a bulk renaming utility and a staff member and I figured out how to use it, but it was not intuitive and the instructions were not clear. I'm told there are other renaming utilities but I didn't investigate further. Once complete, though, I was uncertain what the process of renaming the files would do in terms of downloading them and matching to the accompanying xml files later. My POWRR colleague said the batch

rename can be run in reverse if a log is created, but I did not test to this extent. IWU cannot support a node and so could not participate as a standalone (meaning, a Private LOCKSS Network) member of MA, but the ftp protocol and bandwidth needed to connect to a member-node seem within our reach. MA is a dark archive meant as back up for problems with local copies, so this would mean our long term storage issues on campus would also continue. I think the need to pay for such a backend service while also paying for campus-maintained servers and front-end user access points might be a difficult distinction to convey to my stakeholders.

Since this is the oldest community-supported LOCKSS system we worked with, I thought it would be easy to find out what trends they've noticed in file degradation. I thought this information would help me articulate the need for periodic replacement. But when I asked how many files have had to be replaced I was told that MA had never collected that data and would have to get the members' buy-in to do the programming work needed to analyze such data. I have to say this surprises me. It seems logical that making a case for digital preservation to people who hold the purse strings should include the ability to use risk assessment language like "Over 10 years, professional digital preservation groups have found that X format types fail at X rate or at X intervals after creation." I do not feel comfortable asking for thousands of dollars per year to do DP because a theory says file degradation happens often enough to worry about it!

I transferred about 8GB of content from our IR (Digital Commons, hosted by bepress) and observed the following: about 200 of the 16,000+ files were rejected at the node server. At this point, the POWRR institution serving as our node does not know why.

My IR is not able to identify newly added content and so our only available option is to transfer bulk downloads on a quarterly basis. At this point, I see two problems with this: 1) MA can only detect duplicates if the original "bag" is opened and the "new" content is placed in it for comparison, and 2) I am not sure how compressing and decompressing in bulk would affect the transfer between all the systems involved. I know that MA can select to have uncompressed content transferred from their servers to the node, but I don't know what long-term data integrity implications there are for compression protocols in place between our IR and my institution, then between my institution and the node, and finally the node to MA. There are a lot of opportunities for problems in that sequence, and I'd want to resolve my questions about them before committing to a long term process. There is talk of MA developing a direct harvesting process from DigitalCommons, but I don't know if the distributed node model would be able to accommodate this step on our behalf.

Preservica (http://preservica.com) is a commercial, hosted macroservice preservation product that incorporates all OAIS recommended processing workflows in one place. The workflows are not entirely intuitive but probably wouldn't take long to get used to. They also developed a web-based training program (replacing their 1.5 day site-visit requirement) which will help keep this

cost affordable. The webinar was helpful, but it was difficult to follow the steps on the presenter's screen and simultaneously practice the steps. I used side-by-side browsers open on my monitor and couldn't always scroll to or see the part of the screen being talked about before the presenter moved on.

The tool is robust and their support (both documentation and people) is extensive. They offer two storage systems but they are both Amazon products (S3 and Glacier), and I have reservations about Amazon's lack of transparency. Preservica says they--Amazon--do not disclose server locations or how often they've needed to perform file restorations to date. In just the two year period of this grant, Preservica (formally known as Tessella) has made progress in its own transparency and its pricing options. Even before their recent enhancements, many institutions were rating the product and support highly. With the new features and pricing options, they will have more clients in the near future. However, the product is more elaborate than IWU needs at this time.

Tools not fully tested:

Curators' Workbench (http://blogs.lib.unc.edu/cdr/index.php/about-the-curators-workbench): I did not have any difficulty downloading this open source software, but I never completed testing it because I did not have collections with the required metadata files to practice on. There was no documentation on how to create and store such files or even a suggestion of a minimum amount of data needed. They require metadata be done in MODS and there are no local experts on that standard available. I found some documentation for cross-walks from different metadata types, but this also was difficult for me to follow. Finally, I work better from written step-by-step instructions, so the developer's use of YouTube as the primary documentation source was difficult for me and the image quality was poor when the screen was enlarged. I like the idea of CW but more development of the pre-use instructions (or warnings about knowledge/expertise assumptions being made) needs to be done before I'd be able to use it. The developer told a POWRR member that they've stopped work on it for now, so time invested on the user's part does not seem like it would be well spent.

Hoppla (Home and Office Painless Persistent Long-term Archiving; http://www.ifs.tuwien.ac.at/dp/hoppla) was abandoned by our project as soon as we found out the developers were no longer working on it. I had already tried it, though, and feel its purpose as automated backup and migration processes seemed promising. I was able to download and install without additional help and I was intrigued by the possibilities. The digital curation community and certainly individuals needing personal digital preservation help would be well-served if someone takes up its development again.