Planned Tool Implementations at UMD

This is a description of the extractors and converters that we plan to implement next at the UMD iSchool. These reflect priorities around curating mostly born digital archival collections.

tool	action	input • data type • extensions • mime-type	output format / fields	use case	Effort	notes
Punch Card Interpreter ODR	Extract	• Image • image/*	• card maker • encoding standard (https://en.wikipedia.org/wiki/Punched_card#Standards) • card data	NARA: Software development to create data conversion tools to de-code 80 column punch cards from bitmap images. NARA has ~700,000 cards that comprise the finding aid for Army Morning Reports on microfilm (~130,000 rolls). With this tool a searchable electronic dataset can be economically assembled from scans of punch cards. The dataset would lead to enhanced access to a highly demanded, large and complex collection through real time searching of Units and Date Ranges in the finding aid as a stand-alone database, in AAD, or through the NARA Catalog.	High	Would need to handle correction tape, as per Shannon Bradley. Do we know the card format and encoding of this particular dataset?
Creation date	Extract	• Document .txt, .pdf, .doc • text/plain, application/pdf, application/msword	• content-based creation date • mean date • dates found	Often files that have been moved through archival deposit workflows or have been moved from computer to computer prior to deposit will no longer have good metadata on the creation date of a document. The algorithm would produce a best guess at a document creation date, based on the various dates used in the text.	Low	Concerned this won't work very often... That is certainly a concern for the creation date. An extract of the content dates would still be useful for indexing purposes. Look at Bill Underwood's work.
Provenance				... As a result duplicates and/or revised versions of documents are present within an archive. A Versus signature based on the File2Learn code will allow us to compare documents to see how similar they are based on text, images, and vector graphics. Can be used to potetially reconstruct the order by which the documents were edited.	Medium	Would need to get File2learn code from Rob Kooper. Was in our SVN repo I believe. Take a look at TACC paper on paragraph alignment and clustering.
Access	Convert	WP, WPD	• MS Word	There is no converter to move documents out of these legacy WordPerfect formats. Once they can be converted to Word or to plain text, they will be accessible in many forms. Roughly 4000 documents in CI-BER match this description so far.	Low	We can use an older Win VM to walk these files to Word..

Locations	Extract	▪ GeospatialFeatureData ▪ .kml, .kmz, .shp ▪ `application/vnd.google-earth.kml+xml`, `application/vnd.google-earth.kmz`, `application/octet-stream`	▪ geospatial bounding box	Allows geospatial search and discovery of relevant archives.	Medium	
File Dependencies	Extract	• shp • mdb • ...		Determine relationships/dependencies between files that contain one or more records. For example, it would be nice to know that a particular .dbf file was part of a Shapefile and should not automatically be converted using a tool like SIARD to keep the Shapefile usable. As another example, it would be helpful to know what .mdx or .idx files were associated with a particular .mdb file. Another example would be to know the relationships between all of the files that make up a particular website to ensure you capture everything you want to capture.	Low	
Collection Summary	Extract			Help in getting the big picture across all our holdings. This shows up in places like helping patrons to find content across our holdings that is relevant to their queries, or trying to identify content that should not be released because of national security, personal privacy, etc. Having tools that could summarize/cluster the content without the archivists having to read every single page would be very helpful. If the results could be presented as a visualization or in some other form so that the user could quickly absorb the information that would be even better.	Medium	
XML indexing	Extract	▪ unknown ▪ .xml ▪ text/xml, application/xml	▪ schema ▪ schema-type (DTD/XSD) ▪ is well-formed? ▪ is schema retrievable? ▪ is valid?	Meaningful search over XML files in the archives will hinge on the schema employed. By extracting at the schema we can index it. This is analogous to file characterization within the XML world.	Medium	
Web links	Extract	▪ HyperDocument ▪ .html, .htm ▪ text/html	▪ title ▪ hyperlinks (href, text)	Allows us to create an index of all of the web pages in a web archive of a site for a federal agency, etc.. Text of links can be used to describe the page referenced, becoming additional keywords.	Low	Let's us try pagerank scoring in archives.

Digitized documents	Extract	▪ Scanned Document Image* ▪ image/*	▪ confidence that image is a scanned form ▪ metrics for document layout recogniition	A federal agency will have legacy paper records and often these are scanned into digital form, but rarely become useful as structured data records. Layout recognition metrics, such as the offsets of horizontal and vertical lines, can be used to classify images as depicting a particular type of paper form. These might be tax forms, census forms, or any kind of routine paper record from the pre/post-digital era. Recognition of the layout lets you apply a template that identified document regions for OCR processing.	Medium /High	Would probably depending on the layout of the document, i.e. would need a tempalte for each type? Gregory Jansen is there a type of scanned document we might start with? Sandeep Puthanveetil Satheesan I believe we could throw in the Census forms here (as they are in CIBER). Any others though? Sandeep: Yes, I can only think of those as well. Greg: Any form with boxes would work. I have some other forms from WWII that are interesting to us at UMD.
Map Recognizer	Extract	• Image • image/*	• Identify when an image is a map (historical or modern) • Identify the geographic area depicted	Seems like a very useful tool for building geographic indexes over historical collections or government reports that include maps.	High	I have only the slightest idea of how this would work. Shape recognition on edges of some kind. Recognizing particular areas would require an index of known areas...
Compelling document thumbnails	Extract	• Document • application/pdf	• page thumbnail that includes graphic or photo	Scholarly and mixed use digital repositories often generate document thumbnails that show the front page, which is usually devoid of images and boring. This extractor will generate an image for the most colorful page in a multi-page document, falling back to the front page strategy.	Medium	A colleague implemented this as a one-off algorithm for the Carolina Digital Repository's "peek at the repository" feature: https://cdr.lib.unc.edu/#p The result is much more interesting that the usual thumbnails, for almost any kind of document. Prior art here: https://github.com/UNC-Libraries/peek-data
Access	Convert	• Proprietary Databases • MDB, DB, DBF	• SIARD Software-Independent Archiving of Relational Databases	2378 MDB (MS Access) files in CI-BER with no converter. 2013 DB files (Paradox / XTreeGold / dbvista / Oracle / XoftSpySE). 432802 DBF files. Ensure that we have the means to access the database tables in all the CI-BER collections. Instrumenting SIARD migration will enable a vendor neutral access format and offer advantages for archives implementing pro-active database migration for long term access.	Medium	https://www.loc.gov/preservation/digital/formats/fdd/fdd000426.shtml#specs
Access	Convert	• Adobe Photoshop Images • PSD	• TIFF (or similar)	PSD files are common in born digital archives, but currently they have no conversion path to standard image formats.	Low	Auto Hotkey for a Windows VM? Can we put it in a Docker container somehow?
Redaction of incidental faces	Convert	• image/*		There are plenty of image collection workflows in the sciences and other areas that incidentally collect images of people's faces. This tool would use existing facial detection routines to blur the area where a face is recognized and return a redacted image.	Medium	See human faces extractor.. Standard OpenCV for blur.