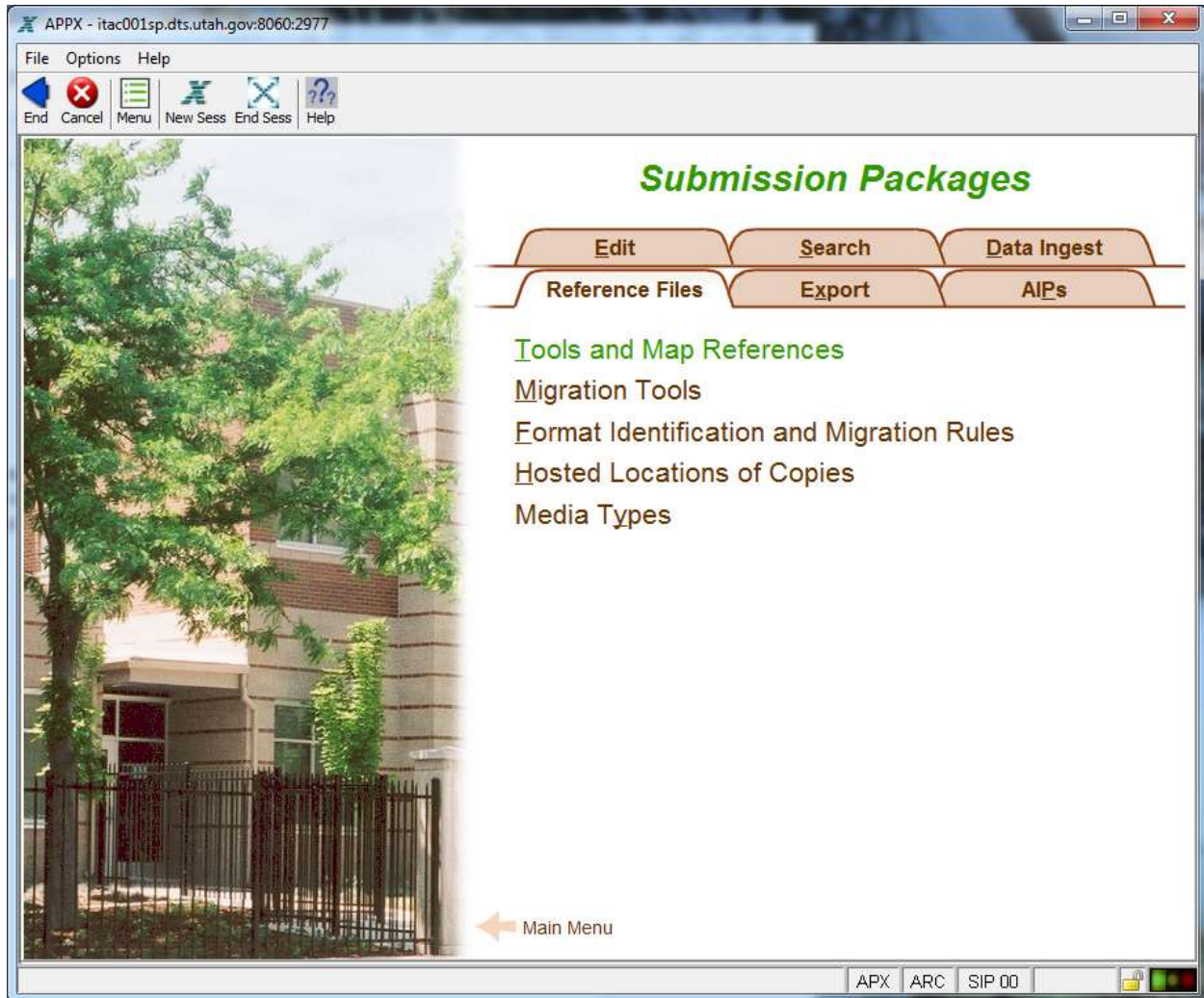# Electronic Reference Files

June 1, 2016

The Reference Files tab on the Electronic Records (AIP) and Submission Packages (SIP) menus includes information about external tools used in the metadata extraction and file migration process. Go to the SIP menu version and click on Tools and Map References.

| Tool Name	Target Files	Version	Map Name	Deflt?	Tool Location
CONTENTDM	Object Groups	1.0.0	SIP-CONTENTDM DUBLIN	✓	
DROID	Electronic Records	6.1.3	NONE		/usr/local/appx/utilities/dro
DROID	Electronic Records	6.1.3 -Nr	NONE	✓	/usr/local/appx/utilities/dro
FGDC	Electronic Records	1.0.0	SIP-FGDC IMPORT	✓	
FGDC	Object Groups	1.0.0	SIP-FGDC IMPORT	✓	
FITS	Electronic Records	0.8.0	SIP-FITS IMPORT	✓	/usr/local/appx/utilities/fit
FITS	Object Groups	0.8.0	SIP-FITS IMPORT	✓	/usr/local/appx/utilities/fit
IMAGE MAGICK	Electronic Records	2010	NONE	✓	/usr/bin/convert
JHOVE2	Electronic Records	2.1.0	SIP-JHOVE IMPORT	✓	/usr/local/appx/utilities/jho
JHOVE2	Object Groups	2.1.0	SIP-JHOVE2 TO OBJGRO	✓	/usr/local/appx/utilities/jho
MEDIAINFO	Electronic Records	0.7.26	SIP-MEDIAINFO	✓	/usr/bin/mediainfo
MEDIAINFO	Object Groups	0.7.26	SIP-MEDIAINFO	✓	/usr/bin/mediainfo
NZ DATA EXTRACT	Electronic Records	3.6	SIP-NZDATA IMPORT	✓	/usr/local/appx/utilities/met

Each tool that is available to the database and produces XML output that needs parsing is listed here, including the version of the tool being used. It's possible to list more than one version of a tool, and make one the default, so users can switch if needed. Generally, though, the newer tools have better options. Note that the Droid option above runs the non-profile version (6.1.3 –Nr) of the extraction process due to its tendency to hang when run from the command line if the full profile version is used. On the desktop, running the full profile will provide additional guesses about a format based on the file extension in addition to signature files, while the non-profile version only references signature files. A signature file is a set of patterns that the software uses to discern what file format is being used. These signature files are published regularly by PRONOM (see http://www.nationalarchives.gov.uk/aboutapps/pronom/droid-signature-files.htm). PRONOM provides a unique identifier for each format called a PUID. This identifier is used by all digital preservation tools in order to speak the same language about format types. During the ingest process, a query will be sent to PRONOM to check if AXAEM is using the latest signature files, and if not, will download them. That delays the ingest process for a minute while the new files are being installed.

The Map Name is the set of instructions used to pull data from XML files into AXAEM fields. Maps between the SIP and AIP versions are similar, but since the data lands in different fields, the maps are distinct. If a tool is updated, both the SIP and AIP versions of the tool name/version/map/location need to be updated.

The Tool Location field identifies where the tool resides on the server, and is important to be accurate so that when AXAEM calls a tool to run, it can find it.

Click Cancel to return to the menu, and then click Migration Tools:
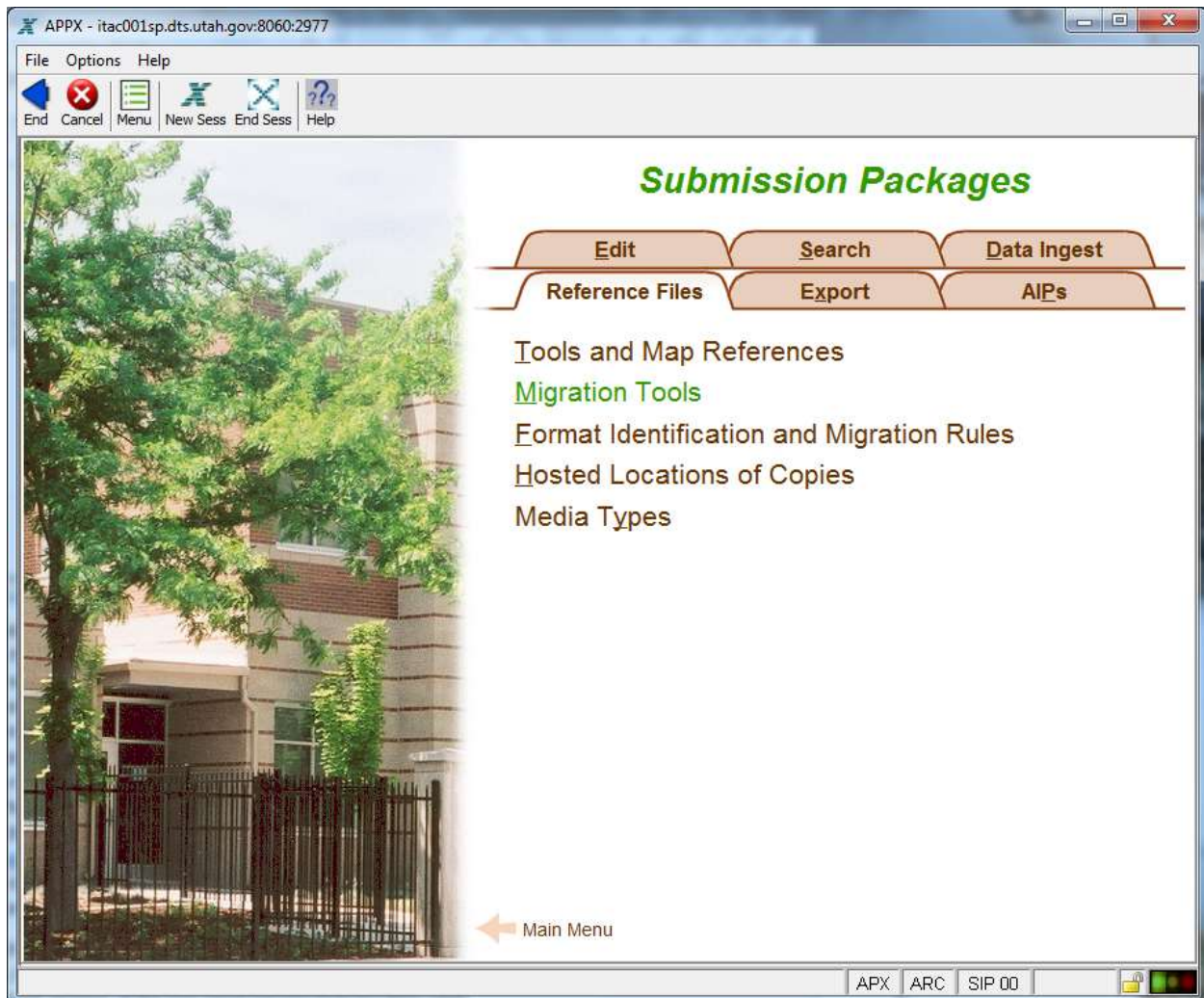

A handful of tools have been added to AXAEM to facilitate the format migration process, which at the moment only really works when creating an access copy during the ingest process. More tools can be added to support more format types as they become available. With the current set of tools, image files can be changed to another image format, and data processing formats can be changed to PDF. Another tool, PDFBox, might be used to change a regular PDF to a PDF/A, but that has not been installed yet. The FFMpeg tool, which is already installed as a metadata extractor for video formats, might also be used to reformat.
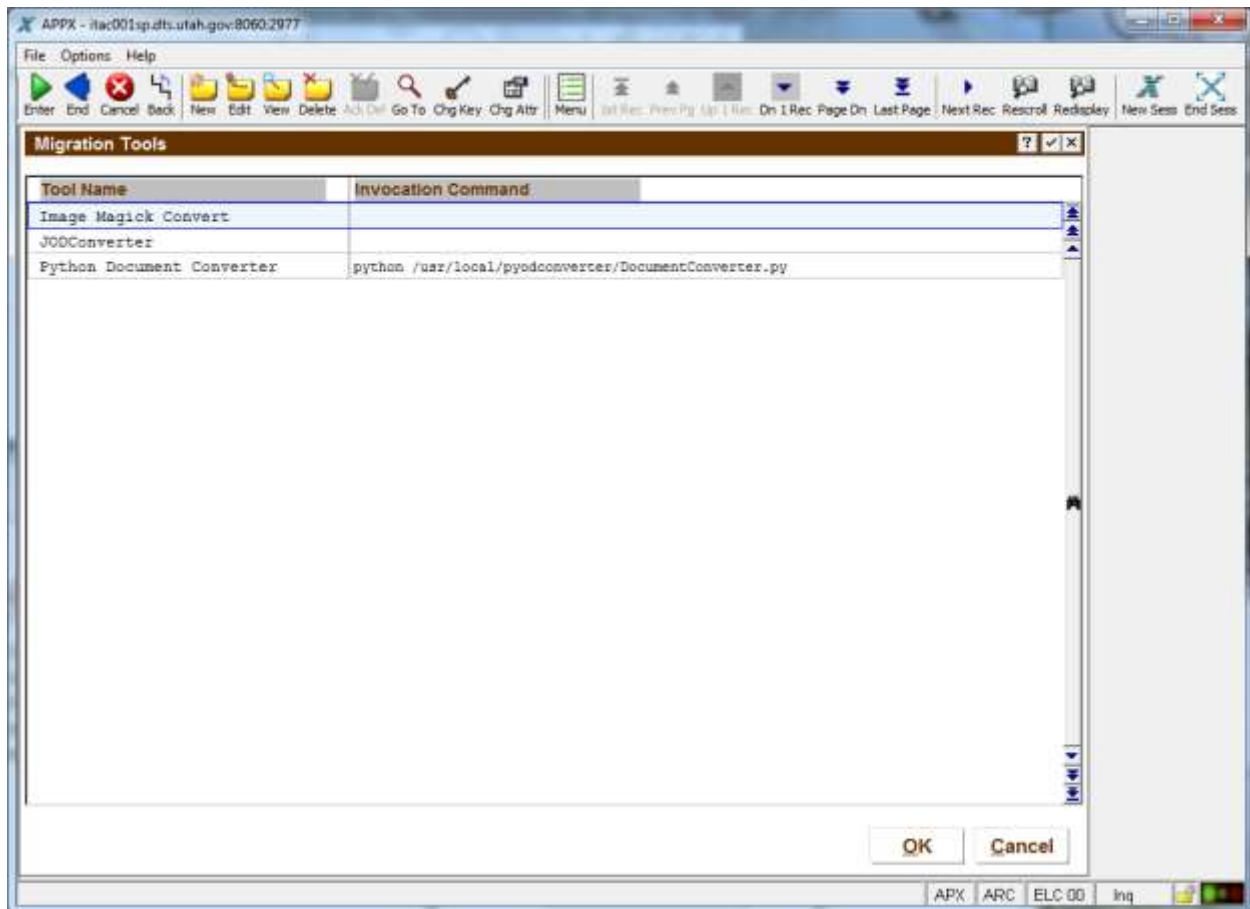
Double-click one of the tools to bring up the formats it supports:
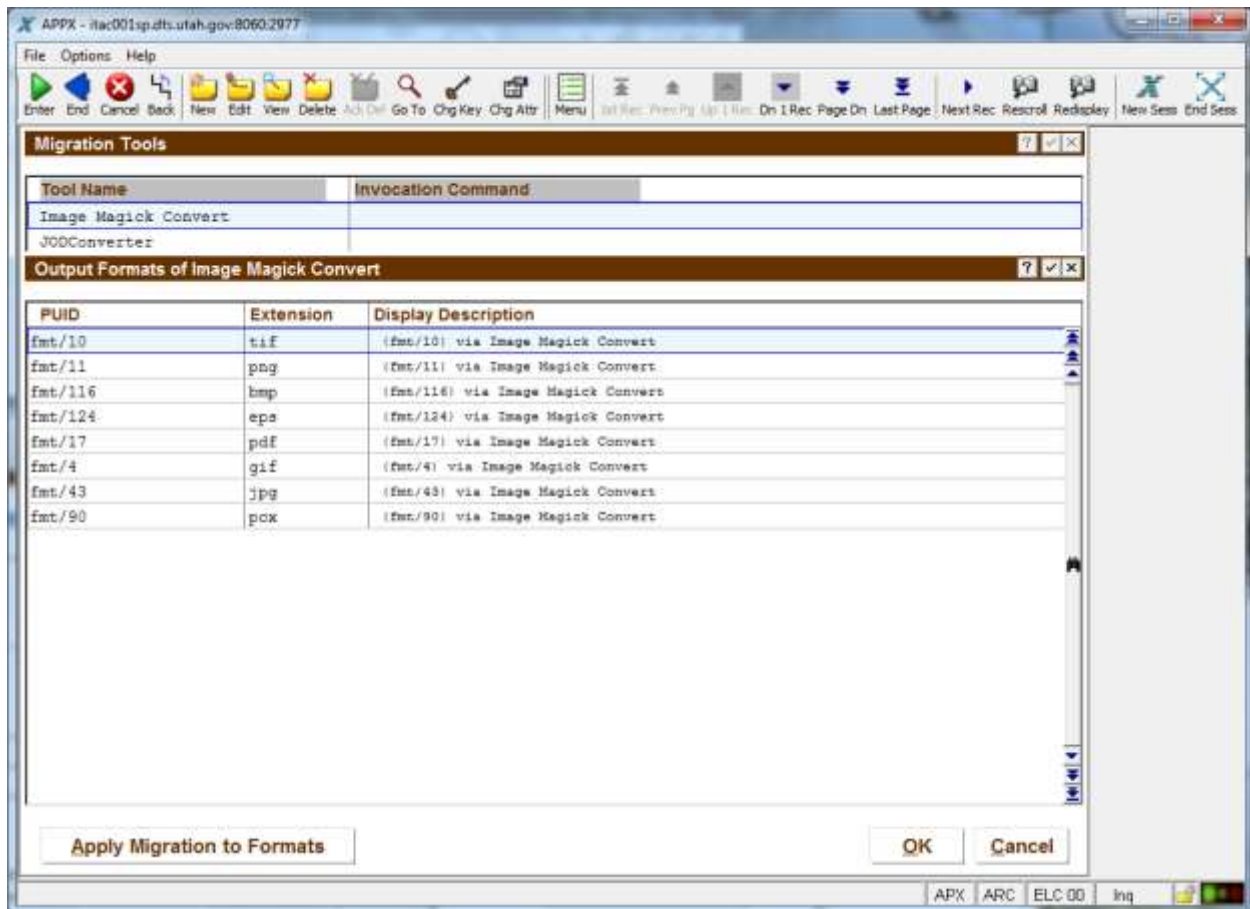
This is where a great deal of data entry needs to happen, so that the system understands exactly which formats the tool supports, so someone will need to research the Image Magick tool, for instance, find a list of all formats it can read and write to, and then compare those formats with the PRONOM identifier (PUID), and add them to the list. The first list is the output format expected. Click OK or double-click a row to drill down to see the source (input) format type:
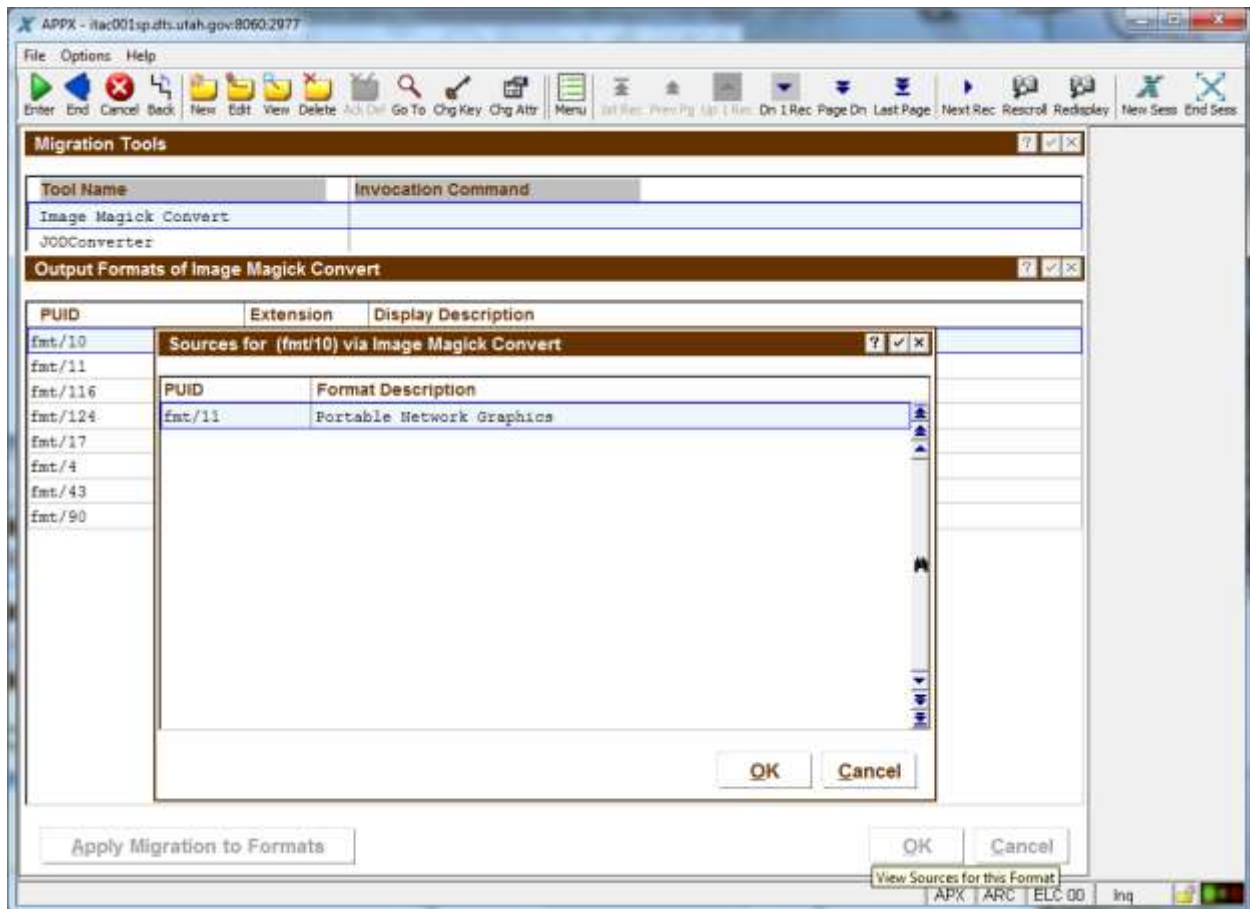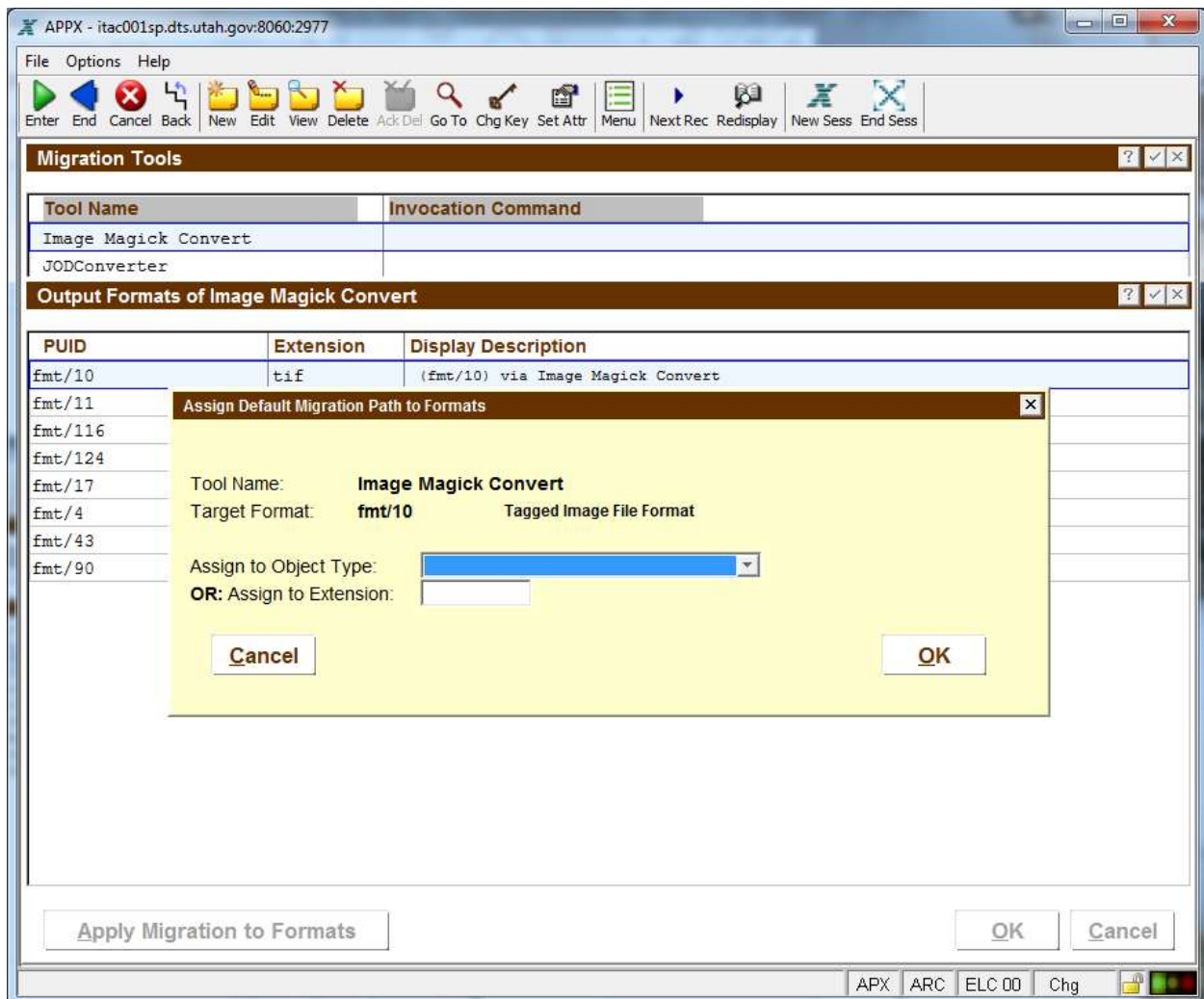
What this effectively means is that every time the system sees a .png file, it can use Image Magick to convert the file to fmt/10 (a .tiff file).

Click the Back button to return to the previous screen. Clicking Cancel will exit you out to the menu.

Click on Apply Migration to Formats:

If the migration process is using a tool on a specific format, metadata here will auto-update the Format definitions with object type information or extensions.

Click Cancel to return to the Reference Files menu. Click Format Identification and Migration Rules:
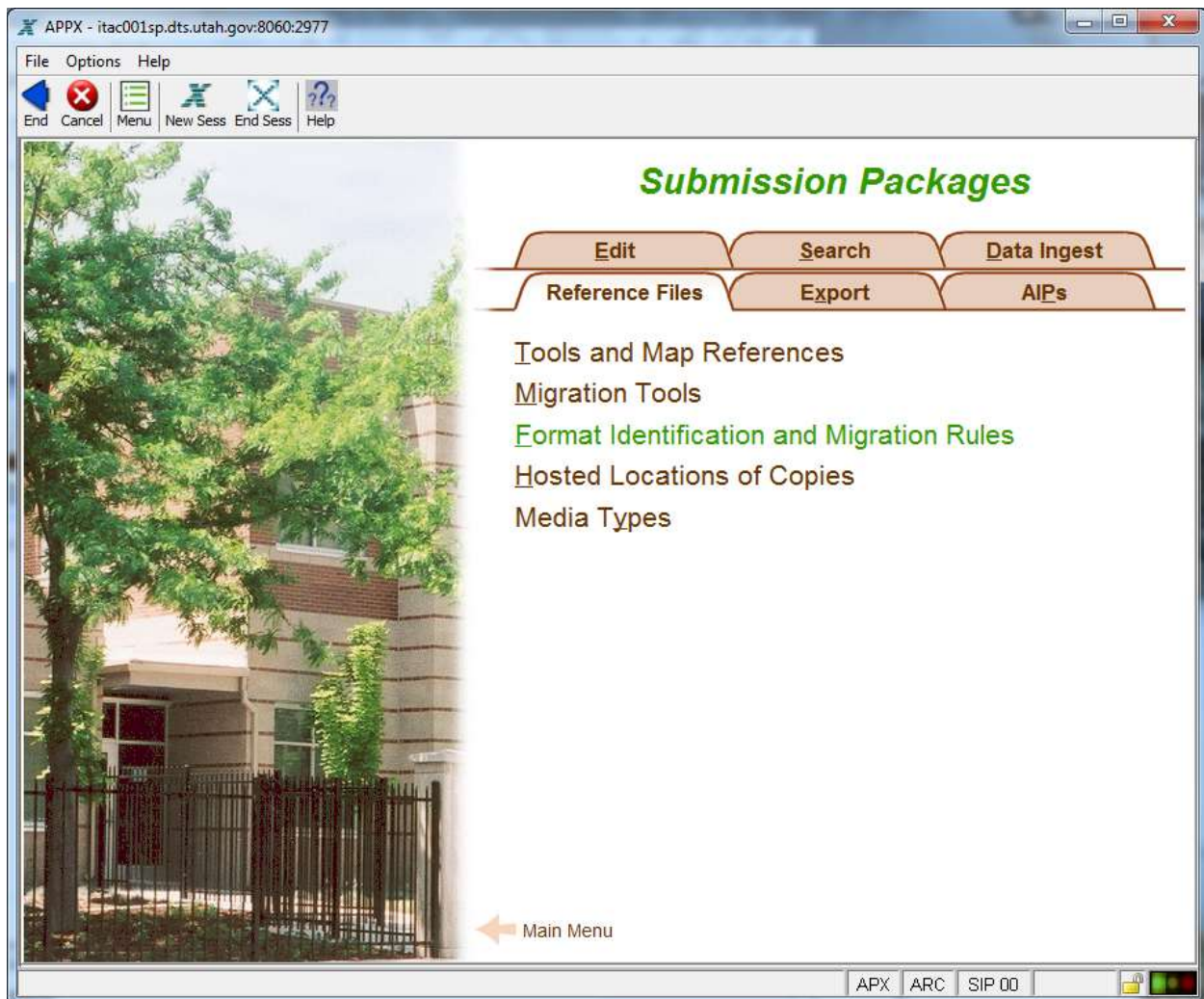
The Format Identification screen provides a table of all the format types currently known by the database. Most of the information was produced from PRONOM signature files, although format types can be added manually if PRONOM is not yet aware of them. Double-click on a row to enter the detail view of the format.

File   Options   Help

Enter End Cancel Back | New Edit View Delete Ack Del Go To Chg Key Set Attr | Menu Next Rec Redisplay | New Sess End Sess

Formats

PUID	Format Name	Object Type	Format Extension(s)	Num	Key
fmt/1	Broadcast WAVE	Audio	.wav	1	1
x-fmt/1	Microsoft Word for Macintosh Document	Word-Processing	Unknown, .mcw	1	2
fmt/2	Broadcast WAVE	Audio	.wav	2	3
x-fmt/2	Microsoft Word for Macintosh Document	Word-Processing	.doc	2	4
fmt/3	Graphics Interchange Format	Image	.gif	3	5
x-fmt/3	Online Description Tool Format	Multi-format	.odt	3	6
fmt/4	Graphics Interchange Format	Image	.gif	4	7
x-fmt/4	Write for Windows Document	Word-Processing	.wri	4	8
fmt/5	Audio/Video Interleaved Format	Video	.avi	5	9
x-fmt/5	Works for Macintosh Document	Multi-format	Unknown	5	10
fmt/6	Waveform Audio	Audio	.wav	6	11
x-fmt/6	FoxPro Database	Database	.dbf	6	12
fmt/7	Tagged Image File Format	Image	.tif, .tiff	7	13
x-fmt/7	FoxPro Database	Database	.dbf	7	14
fmt/8	Tagged Image File Format	Image	.tif, .tiff	8	15
x-fmt/8	dBASE Database	Database	.dbf	8	16
fmt/9	Tagged Image File Format	Image	.tiff, .tiff	9	17
x-fmt/9	dBASE Database	Database	.dbf	9	18
fmt/10	Tagged Image File Format	Image	.tiff, .tiff	10	19
x-fmt/10	dBASE Database	Database	.dbf	10	20
fmt/11	Portable Network Graphics	Image	.png	11	21
x-fmt/11	Revisable-Form-Text Document Content Architecture		Unknown	11	22
fmt/12	Portable Network Graphics	Image	.png	12	23
x-fmt/12	Write for Windows Document	Word-Processing	.wri	12	24
fmt/13	Portable Network Graphics	Image	.png	13	25
x-fmt/13	Tab-separated values	Delimited-Text	.tsv	13	26
fmt/14	Acrobat PDF 1.0 - Portable Document Format	PDF	.pdf	14	27
x-fmt/14	Macintosh Text File	Text		14	28

Detail   **Add New**   **Cancel**

APX | ARC | ELC 00 | Inq

Add additional details you know about the format type, including its history and prevalence. Attach documentation of technical specifications if they are open and available. If policies state that any time this format is found, it should be auto-migrated to a different format, provide the default migration path, plus additional migration paths if available. Click on the Migration Paths option:
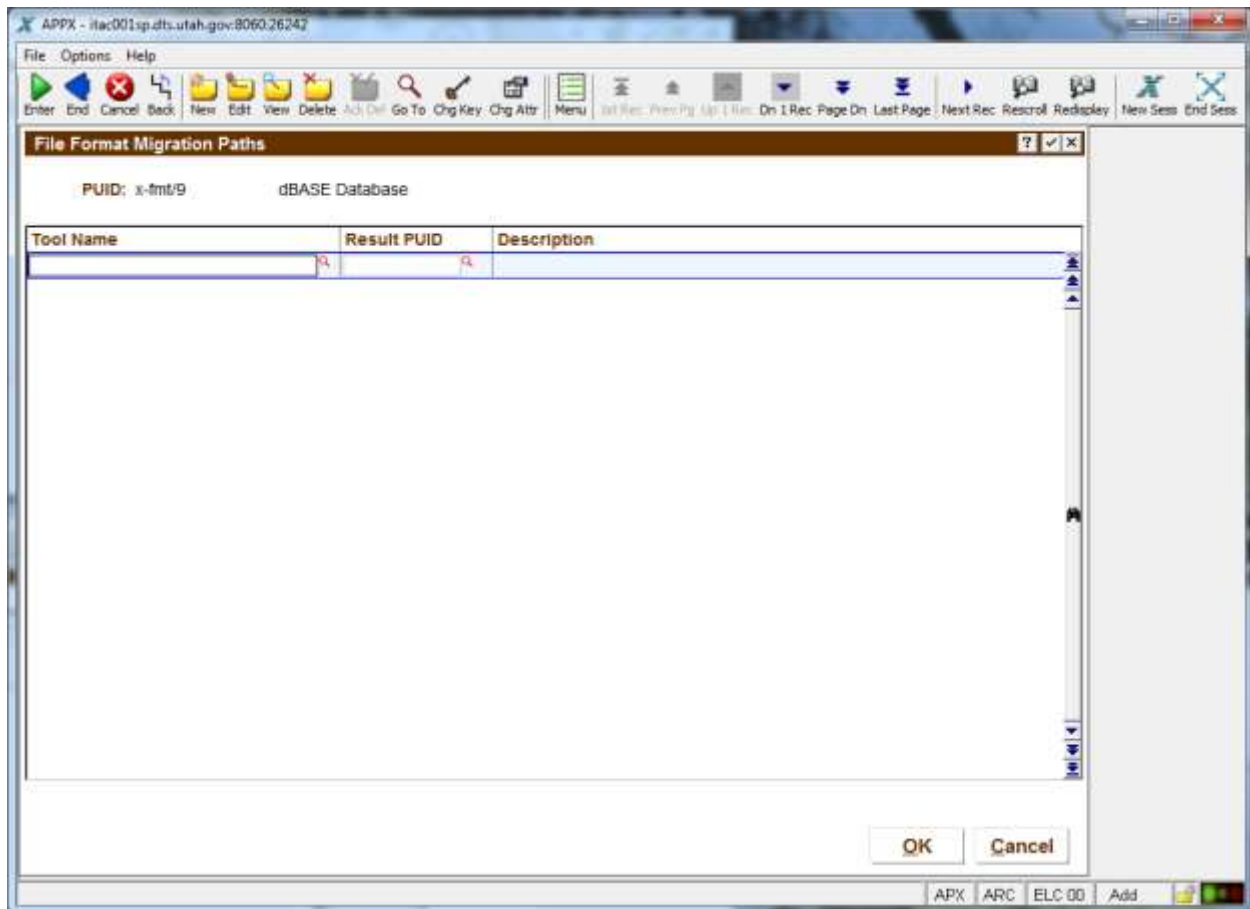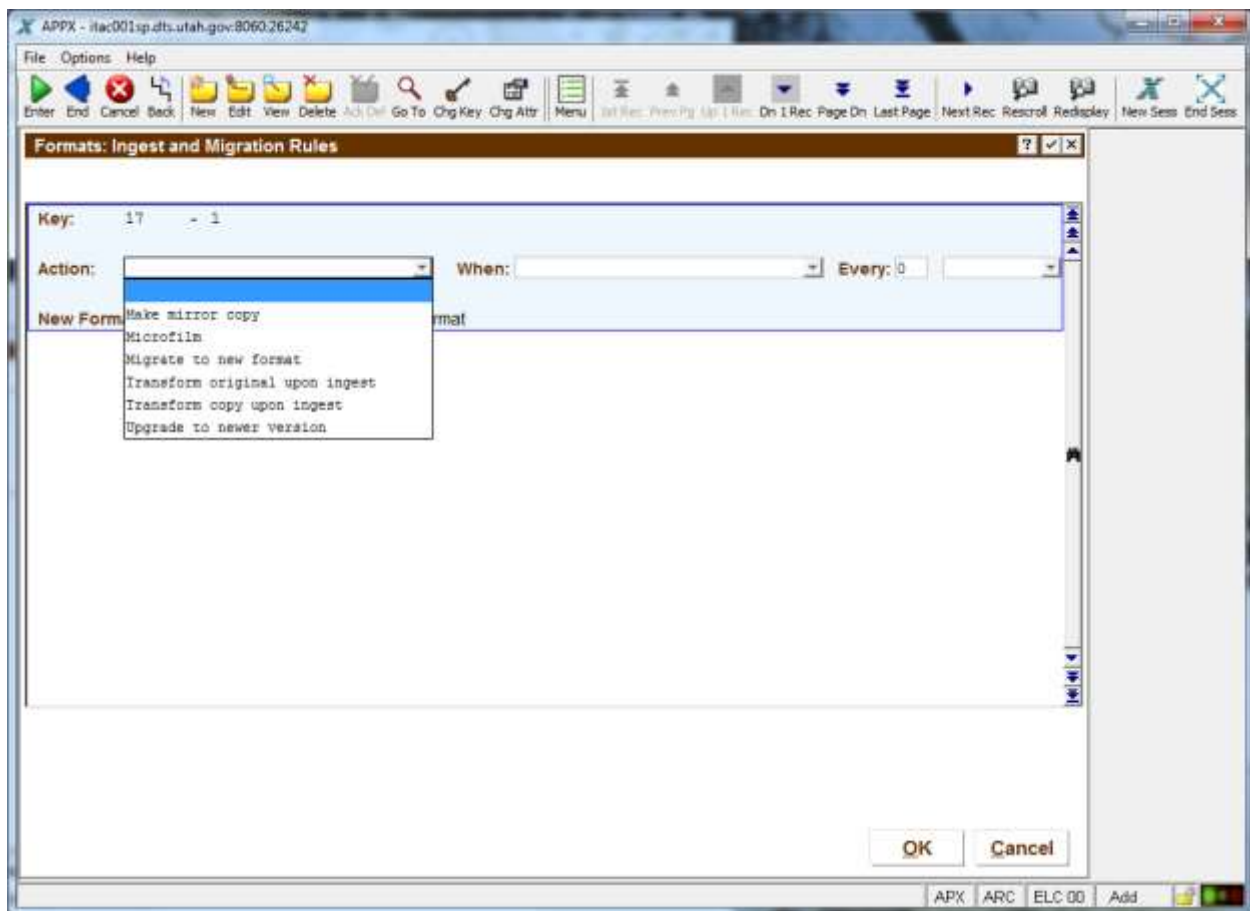
Add the tool name to be used, and the resulting PUID (format identifier). Click OK to save your changes, or Cancel to return to the Formats screen.
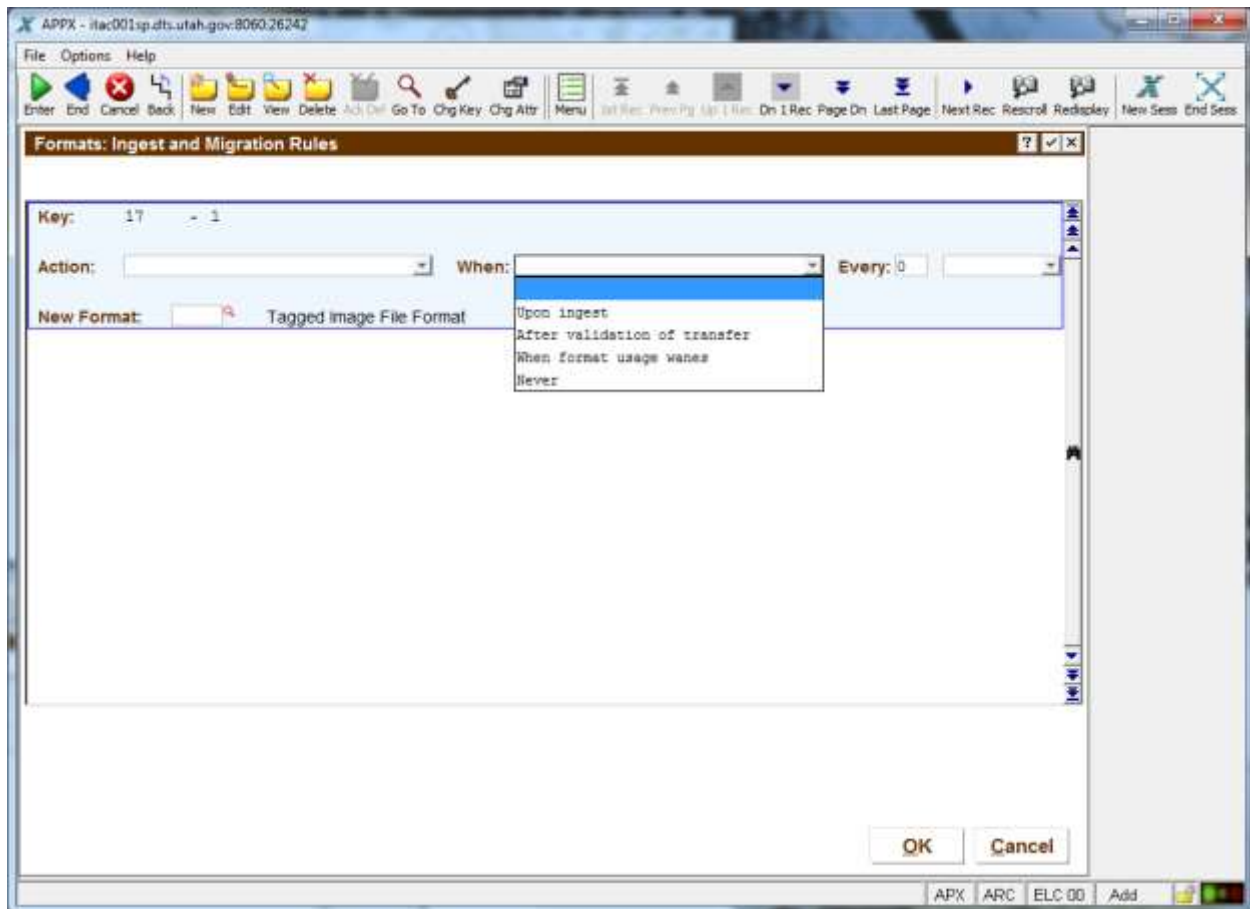
If a format should be represented by a particular thumbnail image for use with the Solr search engine, then indicate that. [Bug report: the GENTHUMB INPUT process is not on a menu that users can access to upload thumbnail images, but when run, looks like this:
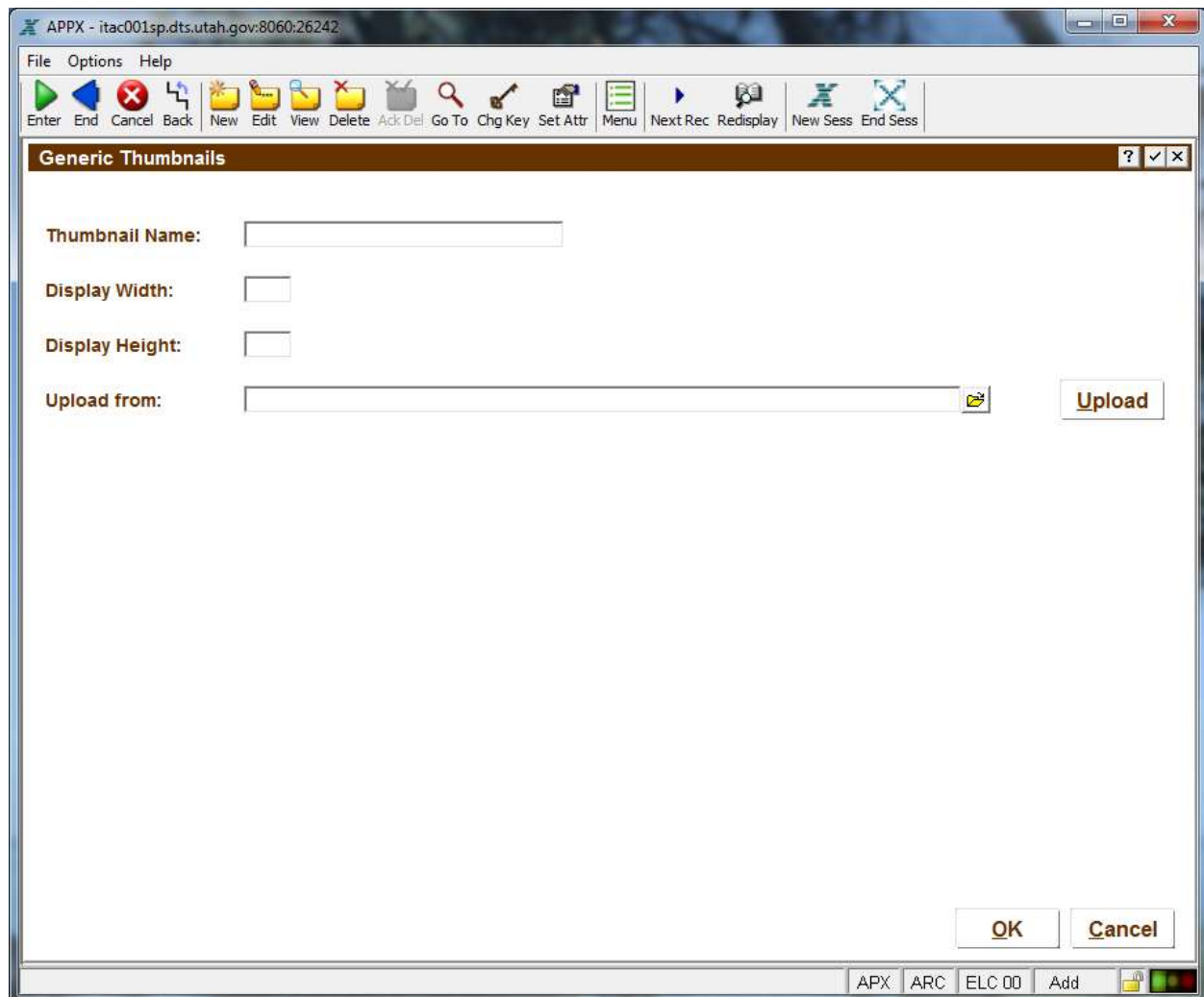
And should be added as an addition to the Reference Files menu.]

The Rules Applied option provides a space to indicate what should happen at various moments in the lifecycle of the format:

File   Options   Help

Enter End Cancel Back New Edit View Delete Ack Del Go To Chg Key Chg Attr Menu 1st Rec Prev Pg Up 1 Rec Dn 1 Rec Page Dn Last Page Next Rec Rescroll Redisplay New Sess End Sess

Formats: Ingest and Migration Rules

Key: 17 - 1

Action: [dropdown]   **When:** [   ]   **Every:** 0   [   ]

New Form Make mirror copy   rmat
Microfilm
Migrate to new format
Transform original upon ingest
Transform copy upon ingest
Upgrade to newer version

OK   Cancel

APX   ARC   ELC 00   Add

Note that these instructions are not connected to any automatic action on the part of the application. They are human-readable policy decisions per format. Some rules could be automated, however. Click OK to save changes or Cancel to return to the Formats screen.

The Extensions option identifies the file extensions commonly used with this format type, as a way of identifying them if signature files do not. The Part of Group checkbox indicates if the format is complex enough to require multiple file types, such as that found in geospatial shapefiles.
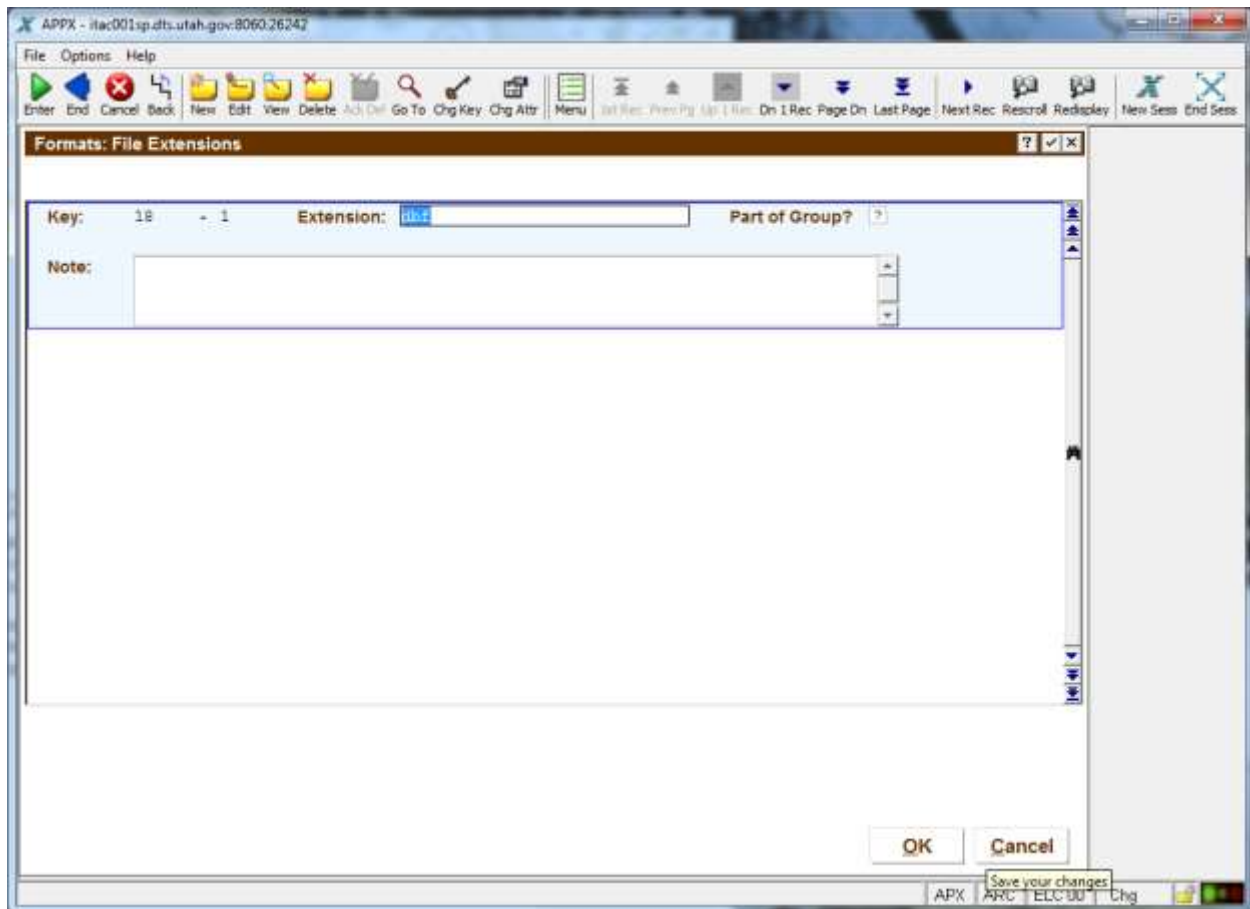
Click OK to save your information or Cancel to return to the Formats screen.

Click on Tools to View Format:

If the software needed to view this format is known, indicate it here. It may be the native software used to create the format, or 3rd-party options which have the ability to view the data.

Click OK to save your changes or Cancel to return to the Formats screen. Click Cancel until you return to the Reference Files menu. Then select Hosted Locations of Copies:
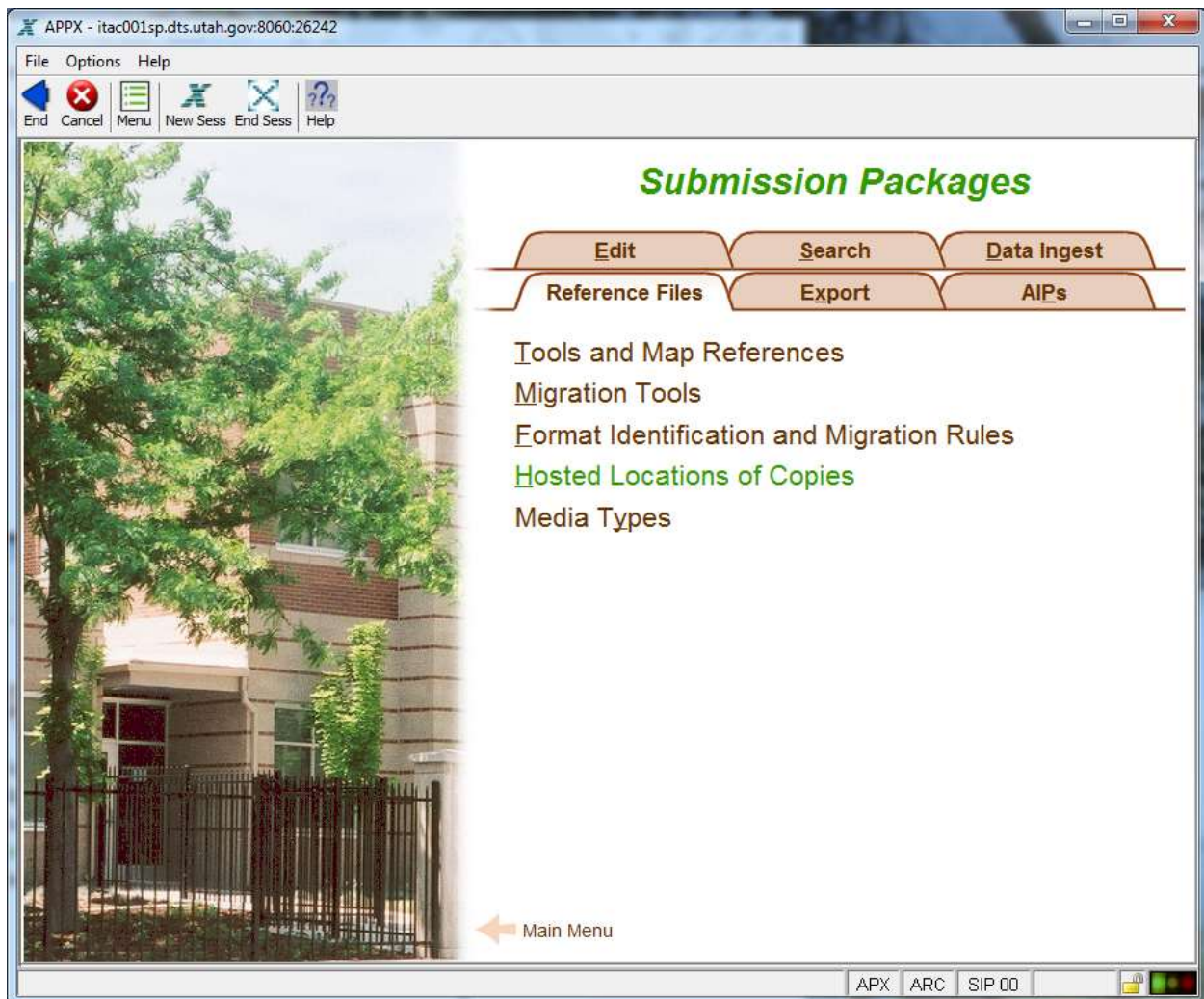
If additional copies of SIPs or AIPs are being hosted remotely by another institution, indicate the location here:

Click OK to save your changes or Cancel to exit back to the menu.

Click on the Media Types option:

File Options Help

End Cancel Menu New Sess End Sess Help

Submission Packages

| Edit | Search | Data Ingest |

| Reference Files | Export | AIPs |

Tools and Map References

Migration Tools

Format Identification and Migration Rules

Hosted Locations of Copies

Media Types

← Main Menu

APX ARC SIP 00

For each media type in use within the electronic records workflow, an identifier should be created here, which will provide instructions about conservation actions and will allow individual media to be linked to a media type. Attach documentation about the media as needed, to help future archivists understand how to operate it, what hardware is required to run it, or other known factors. Click OK to save your changes or Cancel to return to the menu.