Medusa Digital Content Format Registry: What it is and How to Use It

What Is the Digital Content Format Registry?

The Digital Content Format Registry (henceforth referred to as the Registry or Registry) is a publicly accessible reference source for policies and knowledge related to born-digital content stewarded by the University Library at the University of Illinois Urbana-Champaign. The research focus of this tool is to document local knowledge gathered about how to identify and render challenging file formats – particularly formats that present challenges including being associated with a specific version of proprietary software. These formats are often also complex, with dependencies upon hardware, software and other files to render successfully.

The Registry is intended to supplement international file format registry efforts which may contain little to no information on file formats of local interest. The Registry creators have provided a model for a tiered view of file format information, referencing larger and more comprehensive national and international digital preservation and curation efforts (where such efforts exist) with the goal of rendering content for appraisal, preservation and access. However, given the multiplicity of ways in which certain types of files may have been produced, it is unlikely that any reference resource will be able to account for the many varieties that may occur in a particular archives hence our interest in maintaining a knowledgebase informed by local collections.

Entries are not vetted to the extent as those records published in international file format registries. Presently, there is no editorial process in place for vetting information within the registry aside from reviewing and applying the information gathered within a registry entry as a form of quality assessment.

Accessing the Digital Content Format Registry:

- Navigate to Medusa collection registry (you do not have to log in) via this URL: https://medusa.library.illinois.edu/
- From the right side of the interface click on the More link. The following box linking to “More actions” will pop up:

Figure 1 Menu options available from Medusa > More link

The following two options are most relevant to the Registry:

- **Digital Content Format Registry: https://medusa.library.illinois.edu/file_formats**
  This is the landing page for the Registry. You can search/filter for specific file extensions and create a new Digital Content Format entry from this page. List of formats and their logical extensions. The Registry does not reflect the entirety of files currently in Medusa – only those formats for which a format entry was manually created. The development of this came from previous file format research projects

  The Digital Content Formats (XML) produces a list of the format registry entries and their logical extensions in XML formatting.

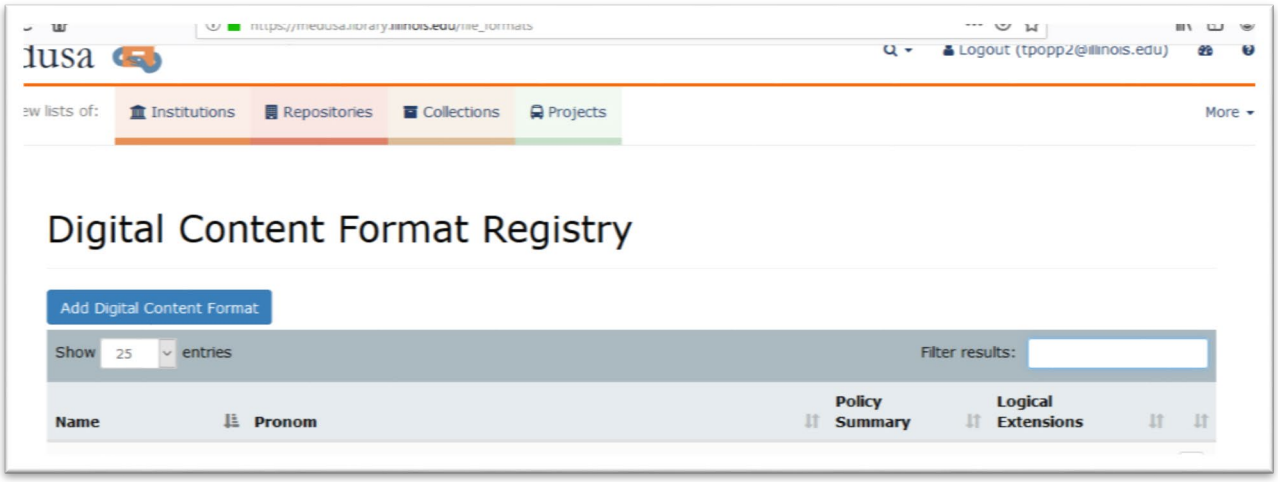


Figure 2 Digital Content Format Registry landing page

- **Rendering Profiles**: Index of rendering profiles associated with file format entries in Medusa. Rendering profiles may include details about opening single files to running complex programs. Provides information about the software and operating system environments used to open files. A profile may also contain a Normalization Path that specifies information about recommended migration target file formats with information about the software recommended for conversion.

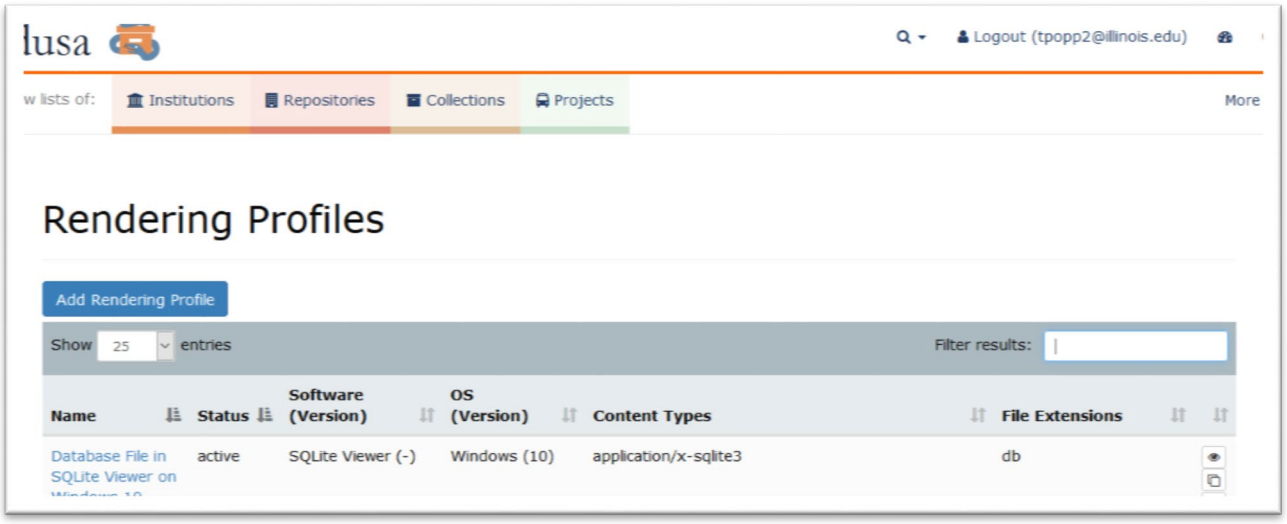


Figure 3 Rendering Profiles Landing Page

Searching for Digital Content Formats

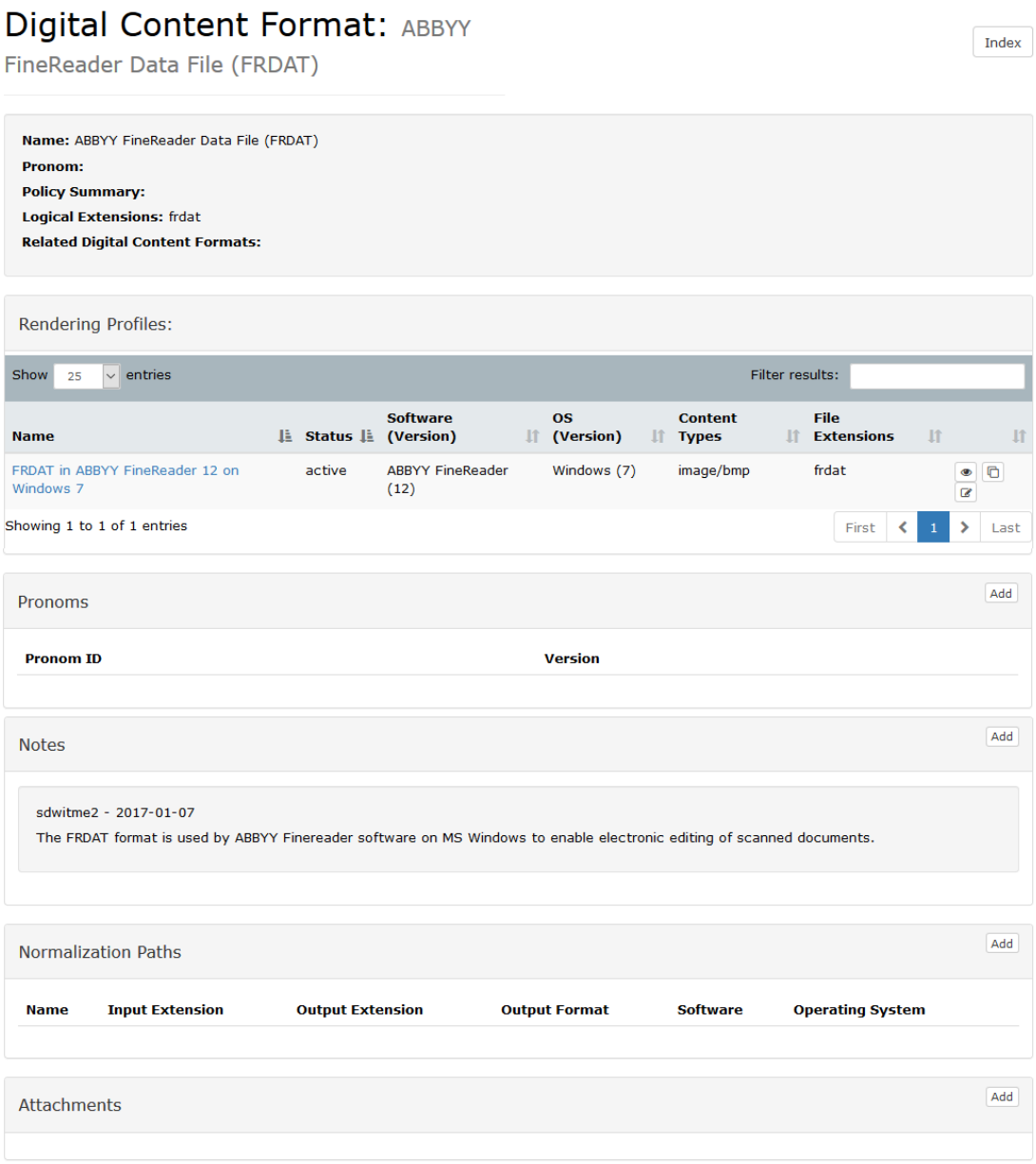
You can search the Registry for entries related to specific attributes by inputting free text in the Filter Results box or by sorting columns alphanumerically. Once you identify a record you would like to review, click on the linked text associated with that record.

Reviewing a Registry Entry

As the Registry is intended to support local needs, each entry may have varying levels of detail depending upon the present ease of accessing a file format or identified priority (or lack thereof) in researching and rendering the file format. Thus, there may be little information about the TIFF file format as at present they are relatively easy to render and do not require significant background knowledge in which to do so. Being a file format standard often used in digitization project, generalities of the TIFF format are often well-represented in other file format registries. Thus, a local entry for this format may be sparse in comparison and refer to local instances and practice.

A complete registry entry is comprised of several parts:

- General Digital Content Format Registry Entry fields
- Rendering Profiles
- PRONOM
- Notes
- Normalization Paths
- Attachments



General Digital Content Format Registry Entry fields

At minimum, a Digital Content Format Registry Entry includes the following fields:

Field	Description
Name	A descriptive, long-form name to identify a particular type of locally held digital content. While these names may simply reference a file format, the concept of a digital content format is flexible enough to encompass variations of file format content profiles, e.g. recurring instances of file formats that share common characteristics.

PRONOM ID	Relevant PRONOM Unique Identifiers (PUIDs) that link to corresponding file format entries within the PRONOM Registry Database. In this model, the Digital Content Format entry is a local enhancement of a base file format entry from PRONOM, in order to document characteristics of digital content formats under institutional stewardship.
Policy Summary	A statement describing local confidence level and curatorial practice with regard to a digital content format, if this exists.
Logical [File] Extension	A list of known file extensions native to the file format or to the program/application from which it originated.
Related Formats	A list of related digital content formats from within the Digital Content Format Registry, some of which may share consistent dependencies.

Rendering Profiles

Rendering profiles may include details about opening single files to running complex programs. They provide information about the software and operating system environments used to open files. The following fields are viewable within the Rendering Profile section:

Field	**Description**
Name	A descriptive, long-form name to identify a particular type of locally held digital content. While these names may simply reference a file format, the concept of a digital content format is flexible enough to encompass variations of file format content profiles, e.g. recurring instances of file formats that share common characteristics.
Status	A rendering profile can either be “active” or “inactive,” which indicates whether a profile is currently (or not currently) being used as a recommended or acceptable path for rendering files.

Software (Version)	Software and version used to open the file.
OS (Version)	Operating system environment in which the software used to render the file was installed and executed.
Content Types	A list of associated MIME types. MIME types are derived from the FITS file characterization tool.
File Extensions	A list of known file extensions native to the file format or to the program/application from which it originated.

PRONOM

Any PRONOM IDs (if applicable) associated with the digital content format. . The following fields are viewable within the PRONOM section:

Field	**Description**
PRONOM ID	Relevant PRONOM Unique Identifiers (PUIDs) that link to corresponding file format entries within the PRONOM Registry Database. In this model, the Digital Content Format entry is a local enhancement of a base file format entry from PRONOM, in order to document characteristics of digital content formats under institutional stewardship.
Version	From the PRONOM registry, copy the associated version information if extant

Notes

Free-text, contextually rich notes and any helpful online and/or local documentation, such as detailed information about relationships and dependencies with other related file formats within the registry.

Normalization Paths

Normalization paths specify information about target file formats to which designated formats should be migrated, with information about recommended software needed for conversion.

Field	**Description**
Name	A descriptive, long-form name to identify a particular type of locally held digital content. The NAME field for the normalization path consist of the following formula (input format + [to] output format + [in] Software + [on] Operating System). For example, the normalization path for the WordPerfect Document entry is named "WordPerfect Document to OpenDocument Text in Xena on Windows 10."
Input logical extension	Source file extension (which can be associated with a multiplicity of formats).
Output logical extension	The target output extension that will be output via the migration process.
Output format	The content format to which the output extension belongs.
Software	The software title and version used in the migration process.

Operating system	OS used to operate the software that migrates a file from one format to another

Attachments

Documents related to the file format or associated software used to render the file, such as published file format standards or software manuals.

Editing Permission:

At this time editing is limited to certain authenticated users within the Medusa system. Note that due to the local nature of the knowledgebase and lack of staff to support editorial control, editing permissions will be limited and will require training. To request editing permissions, send an email providing rationale for requiring editing permissions and an example record you would like to edit to: medusa@library.illinois.edu

References:

For in-depth information about Registry development, refer to the (presently unpublished) paper: Preservation and Access for Born Digital Electronic Records: The Case for an Institutional Digital Content Format Registry (expected publication in Summer 2020. More info TBD. Contact Tracy Popp tpopp2@illinois.edu or Kyle Rimkus rimkus@illinois.edu for more information.)

Document History

Version	Date	Author	Notes
0.1	2019-08-19	Germeck, Karl; Popp, Tracy; Rimkus, Kyle	First draft
1.0	2020-04-21	Germeck, Karl; Popp, Tracy; Rimkus, Kyle	Released for FCoP