DCAT Meeting October 2017

Date & Time

- October 10th **15:00** UTC/GMT - **11:00** ET

Dial-in

We will use the international conference call dial-in. Please follow directions below.

- U.S.A/Canada toll free: 866-740-1260, participant code: 2257295
- International toll free: http://www.readytalk.com/intl
    - Use the above link and input 2257295 and the country you are calling from to get your country's toll-free dial in #
    - Once on the call, enter participant code 2257295

Agenda

Marianne Reed will moderate the October 10th call.
The meeting will focus on a discussion of strategies for determining the levels of support (Supported, Known, Unknown/Unsupported) in the DSpace bitstream format registry and possible implications for preservation.

Preparing for the call

If you can join the call, or are willing to comment on the topics submitted via the meeting page, please add your name, institution, and repository URL to the Call Attendees section below.

Meeting notes

Call focused on what we're doing as a community with the DSpace file format registry.

What are institutions doing with this registry? What do the 'levels of support' mean at your institution - what levels of preservation do these suggest for you? Are institutions modifying the registry that comes packaged with DSpace?

At University of Kansas (KU): as KU was preparing to submit their electronic theses and dissertations to the Digital Preservation Network (DPN), they found a lot of supplemental files that accompanied the primary PDF file. Prompted conversation about these supplemental files - found ~40 filetypes among the set. (Apart from ETDs: Also use GoogleRefine to pull out filetypes from bitstream metadata matched with format registry for items submitted in previous two weeks. Prompts opportunity to intervene in filetypes.)

At University of Edinburgh: use registry essentially out of the box. Find it a bit clunky and suggest it could be made 'smarter' by making information about the formats more readily accessible and searchable. They bring information about the file extensions into a standalone team wiki page to make it more accessible. Will use PRONOM and DROID to automate and track. DSpace **should** be configured to check filename extensions when users upload and showing information about levels of support. Also encountered some messiness with need to record a MIMETYPE for each file format. Have developed largely manual systematic approaches to tracking filetypes for digital preservation and would like to see more of this built into DSpace.

What is the history of this DSpace feature and intentions behind it or any ongoing work to develop it? (No one on the call knew this history.)

Georgetown has added JP2 to their registry but has otherwise not altered.

What are the distinctions between known/unknown/supported? Are institutions backing up 'supported' with official policy? Reed points to MIT's approach as exemplary.

File format registry as an attempt to enable digital preservation but not at the level of robust digital preservation, see Archivematica.

What are institutions doing to migrate file formats? More broadly, how are we restricting or guiding accepted file formats? For data deposit, for example, likely to see a large range of file types, from open to proprietary, preservation-friendly to not. Not seeing file format requirements built into national (in the US) data deposit requirements.

Virginia Tech: Ultimately, they accept everything, all file types, intervene through education and outreach to review file formats and recommend, for example, that someone depositing a PPT also deposit a PDF. Data management services help with this intervention/consultation (though this mostly goes to the data repository rather than DSpace). Graduate School sets policy for ETD supplementary files.

Question of intenvention for theses and dissertations: repository admin tends to be at the end of the line (at the 'tail end of the research lifecycle'), hard to intervene in processes that have generated particular file types for these projects, months ago. Students likely to include the files that they find to be essential to their research, likely too late to be changing or interested in sinking energy in file conversion.

Edinburgh and SAD outreach: emphasize sustainability, accessibility, and discoverability.

Virginia Tech model with data repository (different branding, different interface). Like Toronto, relies on one preservation service in the libraries for multiple systems.

At University of Toronto, support a number of repositories and systems. Questions about preservation component and decision to create single preservation pipeline rather than engineering separate preservation processes/policies for each system. Team is currently constructing this pipeline, which will determine filetype support.

Preparing for future preservation systems and processes.

Is anyone using "supported" to trigger a real migration right now? Who is doing a good job with preservation for DSpace?

How are repository admins auditing their repositories for format? Edinburgh has identified this as a priority in the next six months. Likely to rely on curation tasks that would use DROID plugin and data from PRONOM to identify formats that are going obsolete and to take action. Recommends Digital Preservation Coalition handbook and other resources (can directly request a PDF of the handbook from them). Colleague from Hanover University spoke about commercial auditing systems.

See Terry's comment below on reporting tools in DSpace 6. Mostly used to check for metadata completeness and enhancement.

At Georgetown, ETD workflow flags 'nonconforming' files (very large files, unsupported or unknown file types) with potential for librarian intervention.

What institutions are doing a lot of work in this area? Might look at ETDPlus effort and supporting institutions. Are supplemental files better off in standalone repositories which (might!) have stricter requirements for file formats or curation intervention? Analog precedent of needing to 'migrate' by digitizing supplemental files included as CDs, floppies, etc.

Call Attendees

- Marianne Reed (University of Kansas)
- Claudio Cortese (4Science)
- Mariya Maistrovskaya (University of Toronto)
- Sarah Potvin (Texas A&M University)
- Felicity Dykas (University of Missouri)
- Iryna Kuchma (EIFL)
- Monica Rivero (Rice University)
- Pauline Ward (University of Edinburgh)
- Terrence W Brady (Georgetown)
- Anne Lawrence (Virginia Tech)
- Michele Mennielli (DuraSpace)
- Maureen Walsh (Ohio State University)