Curating Research Data with Hydra and PCDM

Esmé Cowles1
1 Princeton University Library, Princeton, New Jersey, United States, escowles@princeton.edu

CHALLENGES OF MODELING RESEARCH DATA

Beyond the complexities faced by typical asset management or institutional repository systems, research data presents a number of complications, including complex hierarchies of related objects that must be modeled and displayed, a wider array of data formats that must be supported, and domain-specific metadata that is necessary to make data intelligible. In addition, there is greater interest in integrating with external services and vocabularies, such as SHARE Notify[1], EZID[2], and ORCID[3]. Managing these complications often leads to software that is tailored to particular data, making it difficult to maintain or share.

HYDRA FOR MANAGING RESEARCH DATA

Hydra[4] is a framework for building repository applications using Blacklight[5] and Solr[6] for full-featured discovery, and Fedora Commons[7] for storage and preservation. The community of Hydra adopters and developers sustain an ecosystem of tools, documentation, training and events to foster collaboration and bring new adopters into the community. A strength of the Hydra framework is that it empowers adopters to implement their own solutions, customizing all parts of the Hydra stack to meet their local needs. This provides a rich toolset for working with research data, with fully-customizable data models, workflows, indexing, and integration with external systems.

However, the flexibility of Hydra has also been a weakness. Allowing adopters to implement their own approaches locally has resulted in innovation, but has also made it hard to share advances. A number of different technical approaches have been tried to increase code-sharing, but with only limited success. Ultimately, the Hydra applications that are the closest to turn-key applications (Sufia[8] and Avalon[9]) have seen the most widespread adoption. The complexity of research data prevents using these widely-adopted Hydra heads, and exacerbates the broader problem of fragmented tools and unsustainable local development.

HYDRA AND THE PORTLAND COMMON DATA MODEL (PCDM)

PCDM[10] is an effort to make applications and metadata more interoperable by standardizing crucial elements of structure. Defining an intentionally simple data model to make adoption as easy as possible, PCDM provides a baseline for interoperable data. Work towards implementing PCDM in Hydra is currently underway, and has led to a community realignment around collaborative development. Sufia and Worthwhile[11] provided the starting point, and were refactored to provide a more customizable stack that allows adopters to share code more easily, and encourages contributing back to the shared core modules. The Hydra PCDM implementation includes Hydra::Works[12] (a toolkit of repository functionality), Curation Concerns[13] (an application for managing digital objects), and Sufia 7 (the new version of the institutional repository application built atop Curation Concerns).

Hydra::Works and Curation Concerns implement the PCDM Works[14] extension, which provides additional classes for modeling logical structure. For example, books use logical structure to model a table of contents which does not necessarily map cleanly to the physical structure of an item that dictates how it is digitized. In the research data context, logical structure can help organize sets of similar items, such as data files that represent a set of permutations, or visualizations that represent various axes.

Curation Concerns promises to address the challenges of modeling research data, making it easier to model complex data objects, and providing a framework for supporting more file formats and domain-specific metadata. It features custom models and workflows, which can be bundled and shared, with work underway to develop shared modules for common types, such as books and geodata. Sustained commitment to shared development of the core modules lessens the burden on individual adopters and helps encourage collaboration.

REFERENCES


ABOUT THE AUTHOR

Esmé Cowles is a developer at Princeton University Library, and a committer for the Hydra and Fedora projects. He was previously at UC San Diego, working on the UCSD DAMS repository. He has focused on metadata interoperability in particular, including work on the Portland Common Data Model and VRA Core 4.