Contribution ID: **35** Type: **Presentation**

A side serve of metadata: evolving AARNet's tools for data packaging

Monday 28 January 2019 17:40 (20 minutes)

A data package is simply the combination of a dataset with metadata that describes the dataset. The purpose of the data package is to provide sufficient contextual data to make the dataset usable to others. It therefore becomes the basis of a loosely coupled data management platform in which the information is in the dataset, not in the platform. A receiving platform just needs to be able to access and interpret the package.

But of course this is where the complexity, and our story, begins. At AARNet we partnered with Intersect Australia and Western Sydney University to implement a packaging plug-in tool called Collections [1], an extension of a tool called cr8it [2], an ownCloud plugin, which in turn was developed on top of the BagIt [3] file manifest specification. Our goal was simple, give the researcher a tool to support data sharing and repository deposits, basically to assist with data publication (including open data publication). So we made it available through our CloudStor [4] console, where the data is visible to the user.

This talk will discuss the next steps in research data packaging. As our CloudStor service has become ubiquitous amongst Australian researchers, the need has grown for a tool that supports not only sharing, but open data publishing and archiving of data, placing data management and curation at the front of the Research Data Management Lifecycle.

For our next iteration of the Collections tool we are considering the work of two other Australian initiatives. The first is the State Library of New South Wales use of BagIt in their PanDA [5] development for ingest as part of the archiving workflow. The second is the University of Technology Sydney on a new data packaging specification that also builds on the BagIt packaging specification and makes data easier to disseminate and consume. DataCrate [6] formats its machine-readable metadata in JSON-LD and follows the schema.org vocabulary, making data packages instantly consumable for semantic driven workflows and instantly consumable and indexable for discovery platforms. DataCrate also creates rich human-readable metadata in the form of web pages that also describe the "who, what, where"metadata that helps to make user understanding of the data and its provenance.

By refining our packaging plugin AARNet can make our CloudStor service interoperable with institutional repositories and digital archives, or maybe, just maybe, make CloudStor the repository and the archive.

[1] https://support.aarnet.edu.au/hc/en-us/sections/115000264274-CloudStor-Collections
[2] https://github.com/IntersectAustralia/cr8it_doc
[3] https://en.wikipedia.org/wiki/BagIt
[4] https://support.aarnet.edu.au/hc/en-us/categories/200217608-CloudStor
[5] http://www.sl.nsw.gov.au/blogs/panda-digital-asset-ingestion-scale
[6] https://github.com/UTS-eResearch/datacrate

Authors: KENNEDY, Gavin (AARNet); MASON, Ingrid (AARNet); Dr SEFTON, Peter (University of Technology Sydney)

Presenter: KENNEDY, Gavin (AARNet)

Session Classification: Sharing and Collaborative Platforms

Track Classification: Open Data Ecosystems and CS3