Introducing next year's model, the data-crate; applied standards for data-set packaging

Peter Sefton1, Peter Bugeia2

1University of Western Sydney, NSW, Australia, p.sefton@uws.edu.au
2Intersect Australia Ltd, Sydney, Australia, peter.bugeia@intersect.org.au

ABSTRACT

In this paper we look at current options available for storing research data to maximize potential reuse and discoverability, both at the level of data files, and sets of data files, and describe some original work bringing together existing standards and metadata schemas to make well-described, reusable data sets that can be distributed as single files, dubbed "crates" with as much context and provenance as possible. We look at some of the issues in choosing file formats in which to archive and disseminate data, and discuss techniques for adding contextual information which is both human-readable and machine-readable in the context of both institutional and discipline data management practice.

INTRODUCTION

We have reached the time when there is a genuine need to be able to match-up data from different sources; infrastructure projects funded by the Australian National Data Service (ANDS) (4) are now feeding human-readable metadata descriptions to the Research Data Australia (RDA) website. But which standards to use? As Tanenbaum said, "The nice thing about standards is that you have so many to choose from. Furthermore, if you do not like any of them, you can just wait for next year's model" (1). However, when it comes to choosing file format standards for research data, we have found that while there might be many standards there is no single standard for general-purpose research data packaging. It is, however possible to stitch-together a number of different standards to do a reasonable job of packaging and describing research data for archiving and reuse.

There are several issues with standards at the file level. For example consider one of the most commonly supported formats: CSV – or Comma Separated Values. CSV file is actually a non-standard, ie there is no agreed CSV specification, only a set of unreliable conventions used by different software, RFC 4180 (2) notwithstanding. While a CSV file has column headers, there is no way to standardise their meaning. Moving up the complexity chain, the Microsoft Excel based .xslx format is a standard, as is the Open Document Format for spreadsheets but again, even though you can point to a header-row in a spreadsheet and say "that's the header" there is no standard way to label variables in a way that will match with the labels used by other researchers, or to allow discovery of the same kind of data points in hetrogenous data sets. There is a well established standard which does allow for "self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data", NetCDF (3) – we will consider how this might be more broadly adopted in eResearch contexts.

PRINCIPLES FOR DATA PACKAGING

The aim in packaging data into 'sets' for archiving it to maximize its future value. To do this, it is desirable to include as much information as possible about the data. What is it? How was it collected? By whom? How does one interpret variables in tables? What is depicted in photos and diagrams? What does it depend on? In answering these questions we applied the following design principles.

Where possible:

- Use file formats which allow in-board/embedded metadata

- The packaging format should not be data-format-sensitive
- The packaging format should not be research domain specific
- The packaging format should not be technology or platform specific
- The data package should contain as much contextual information as possible
- Metadata should be easily human and machine-readable
- The package format should contain self-checking and verification features

INTRODUCING THE CRATE

When the eResearch team at UWS and Intersect NSW were working on the ANDS DC21 "HIEv" (5) application to allow researchers to create data-sets from collections of files, we looked in vain for a simple-to-implement solution for making CSV-type data available with as much provenance and re-use metadata as possible, as per the principles outlined above. In this presentation we will discuss some of the many file-packaging options which were considered and rejected including METS (6), and plain-old zip files with no metadata. The project devised a new proof-of-concept specification, known as a 'crate', based on a number of standards,. This format:

- Uses the California Digital Libraries Bagit specification(7) for bundling files together into a bag.
- Creates a single-file for the bag using zip (other contenders would include TAR or disk image formats but zip is widely supported across operating systems and software libraries).
- Uses a human-readable HTML README file to make apparent as much metadata as is available from (a) within files and (b) about the context of the research data.
- Uses RDF with the W3C's DCAT ontology (8) and others to add machine readable metadata about the package including relationships between files, technical metadata such as types and sizes and research context information such as which facility or experiment the data relates to.

REFERENCES

1. Tanenbaum AS. Computer networks. Prentice H all PTR (ECS Professional). 1988;1(99):6.
2. <ietf@shaftek.org> YS. Common Format and MIME Type for Comma-Separated Values (CSV) Files [Internet]. [cited 2013 Jun 8]. Available from: http://tools.ietf.org/html/rfc4180
3. Rew R, Davis G. NetCDF: an interface for scientific data access. Computer Graphics and Applications, IEEE. 1990;10(4):76–82.
4. Sandland R. Introduction to ANDS [Internet]. ANDS; 2009. Available from: http://ands.org.au/newsletters/newsletter-2009-07.pdf
5. Intersect. Data Capture for Climate Change and Energy Research: HIEv (AKA DC21) [Internet]. Sydney, Australia; 2013. Available from: http://eresearch.uws.edu.au/blog/projects/data-capture-for-climate-change-and-energy-research/
6. Pearce J, Pearson D, Williams M, Yeadon S. The Australian METS Profile–A Journey about Metadata. D-Lib Magazine. 2008;14(3/4):1082–9873.
7. Kunze J, Boyko A, Vargas B, Madden L, Littman J. The BagIt File Packaging Format (V0.97) [Internet]. [cited 2013 Mar 1]. Available from: http://tools.ietf.org/html/draft-kunze-bagit-06
8. Maali F, Erickson J, Archer P. Data Catalog Vocabulary (DCAT) [Internet]. World Wide Web Consortium; Available from: http://www.w3.org/TR/vocab-dcat/

ABOUT THE AUTHOR(S)

Peter Sefton is the Manager for eResearch at the University of Western Sydney (UWS). Before that he ran the Software Research and development Laboratory at the Australian Digital Futures Institute at the University of Southern Queensland. Following a PhD in computational linguistics in the mid-nineties he has gained extensive experience in the higher education sector in leading the development of IT and business systems to support both learning and research While at USQ, he oversaw the creation of one of the core pieces of research data management infrastructure to be funded by the Australian National Data Service consulting widely with libraries, IT, research offices and eResearch departments at a variety of institutions in the process. The resulting Open Source research data catalogue application ReDBOX is now being widely deployed at Australian universitiesAt UWS Peter is leading a team which is working with key stakeholders to implement university-wide eResearch infrastructure, including an institutional data repository, as well as collaborating widely with research communities at the institution on specific research challenges. His research interests include repositories, digital libraries, and the use of The Web in scholarly communication.

Peter Bugeia is the Intersect eResearch Analyst for the University of Western Sydney. Peter has 30 years IT experience across a wide range of industries including medicine, banking and finance, media, and education. He has worked in commercial, not-for-profit and public sectors and has held various roles from Senior Software Engineer and Test Manager to Project Manager, Enterprise Architect and Business Analyst.