Presentation Proposal: 4A Data Management. Acquiring, Acting-on, Archiving & Advertising research data at the University of Western Sydney

Peter Sefton* p.sefton@uws.edu.au
Peter Bugeia** peter.bugeia@intersect.org.au
*University of Western Sydney
**Intersect Australia Ltd

Abstract

There has been significant Government investment in Australia in repository and eResearch infrastructure over the last several years, to provide all universities with an institutional repository for publications, and via the Australian National Data Service to encourage the creation of institution-wide Research Data Catalogues, and research Data Capture applications. Further rounds of funding have added physical data storage and cloud computing services. This presentation looks at an example of how these streams of money have been channeled together at the University of Western Sydney to create a joined-up vision for research data management across the institution and beyond, creating an environment where data may be used by research teams within and outside of the institution. Alongside of the technical services, we report on early work with researchers to create a culture of replicable use of data, towards the vision of truly reproducible research.

This presentation will show a proven end-to-end design for research data flows, starting from a research group, The Hawkesbury Institute for the Environment, where a large sensor network gathers data for use by institute researchers, in-situ, with data flowing-through to an institutional data repository and catalogue, and thence to Research Data Australia - a national data search engine. We also discuss a parallel workflow with a more generic focus - available to any researcher. We also report on work we have done to improve metadata capture at source, and to create infrastructure that will support the entire research data lifecycle. We include demonstrations of two innovations which have emerged from the associated project work: the first is of a new tool for researchers to find, organize, package and publish datasets; the second is of a new packaging format which has both human-readable and machine-readable components.

The 4A Vision

For the purposes of this presentation we will talk about the '4A' approach to research data management - Acquire, Act, Archive and Advertise. The choice of different terms from the 2Rs Reuse and Reproduce of the conference theme is intended to throw a slightly different light on the same set of issues. The presentation will examine each of these 'A's in turn and explain how they have helped us to organize our thinking in developing a target technical data architecture and integrated data-related end-to-end business processes and services involving research technicians and support staff, researchers and their collaborators, library staff, information technology staff, office of research services, and external service providers such as the Australian National Data Service and the National Library of Australia. The presentation will also discuss how all of this relates to the research project life cycle and grant funding approval.

Acquiring the data

We are attacking data acquisition (known as Data Capture by the Australian National Data Service, ANDS [1]) in two ways:

With discipline specific applications for key research groups. A number of these have been developed in Australia recently (for example MyTARDIS [2]), we will talk about one developed at UWS. With ANDS funding, UWS is building an open source automated research data capture system (the HIEv) for the Hawkesbury Institute for the Environment to automatically gather time-series sensor data and other data from a number of field facilities and experiments, providing researchers and their authorised collaborators with easy self-service discovery and access to that data.
Generic services for Data storage via simple file shares, Integration with cloud storage including Dropbox.com and other distributed file systems. And Source-code repositories such as public and private github and bitbucket stores for working code and textual data.

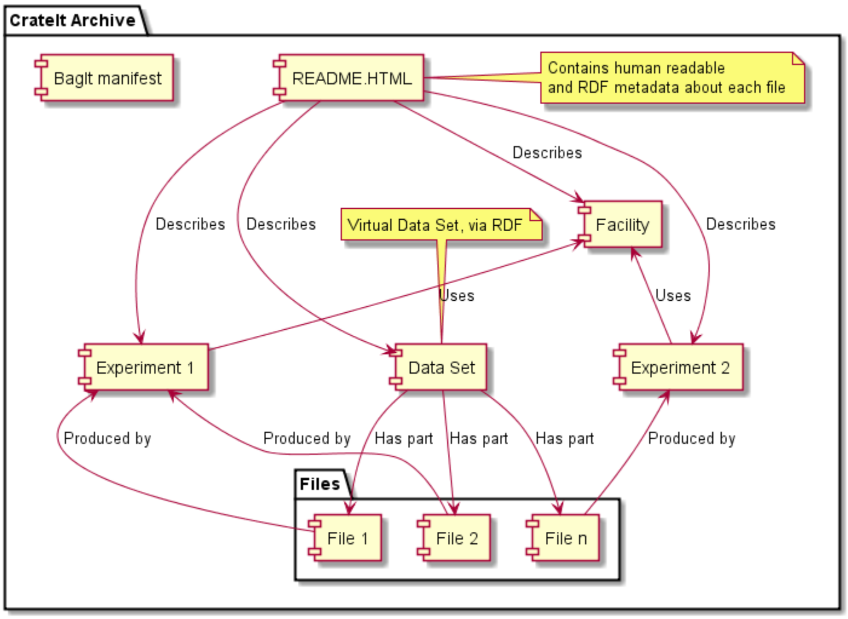
Acting on data

The data Acquisition services described above are there in the first instance to allow researchers to *use* data. With our environmental researchers, we are developing techniques for developing reusable data sets which include raw data, commented scripts to clean the data (eg a comment "filter out known bad-days when the facility was not operating") then re-organize it via resampling or other operations into useful 'clean' data that can be fed to models, plotted etc and used as the basis of publications. Demo: the presentation will include a live demonstration of using HIEv to work on data and create a data archive.

From action to archive

Having created both re-usable base data sets and publication-specific operations on data to create plots etc there are several workflows where various parties trigger deposit of finished, fixed, citable data into a repository. Our project team mapped out several scenarios where data are deposited with different actors and drivers including motivations that are both carrot (my data set will be cited) and stick (the funder/journal says I have to deposit). Services are being crafted to fit in with these identified workflows rather than build new things and assume "they will come".

BagIt structure in a ZIP Archive, can be unzipped onto web server and README.HTML describes the data-set within



Archiving the data

The University of Western Sydney has established a Research Data Repository[i] (RDR), the central component of which is a Research Data Catalogue, running on the ReDBOX open source repository platform. While individual data acquisition applications such as HIEv are considered to have a finite lifespan, the RDR will provide on-going curation of important research datasets. This service is set up to harvest data sets from the working-data applications, including the HIEv data-acquisition application and the CrateIt data packaging service using the Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH).

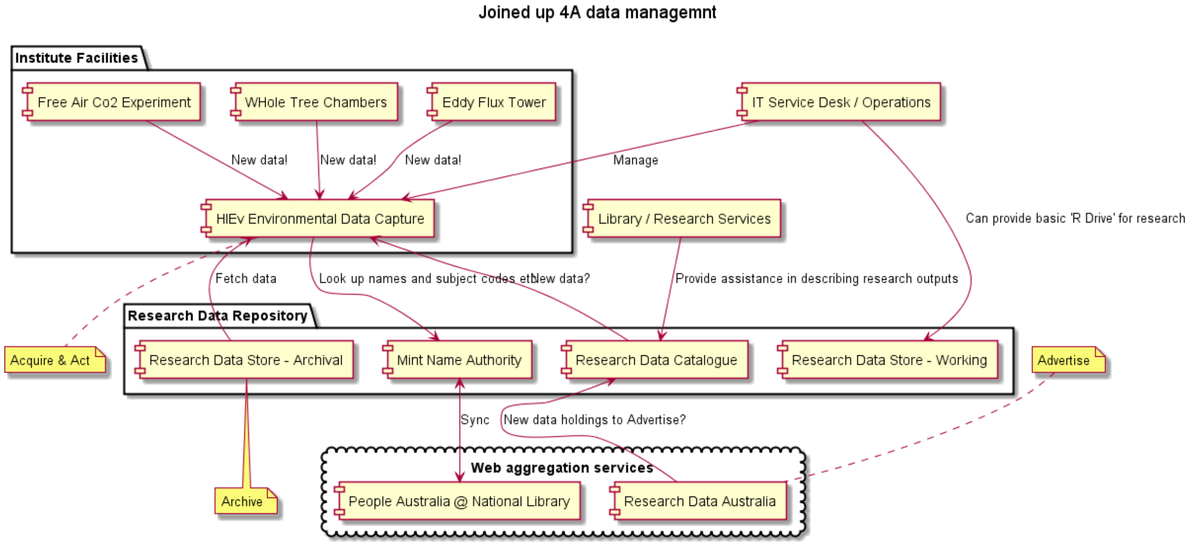
Advertising the data

As with Institutional Publications Repositories, one of the key functions of the Research Data Repository is to disseminate metadata about holdings to aggregation services and give data a web presence. Many Australian institutions are connected to the Research Data Australia discovery service [6], which harvests metadata via an ANDS-defined standard over the OAI-PMH harvesting protocol. There is so far no Google-Scholar-like service which is harvesting data about data sets via direct web crawling (that we know about), so there are no firm standards for how to embed data in a page, but we are tracking the developments of the Schema.org vocabulary, which is driven largely by Google's group of companies which are Google's peers, and the work described above on data packaging with RDFa metadata is intended to be consumed by direct crawlers. It is possible to unzip a CrateIt package and expose it to the web thus creating a machine-readable entry-point to the data within the Zip/BagIt archive.

Looking to the future, the University is also considering plans for an over-arching discovery hub, which would bring together all metadata data about research including information on publications, people, and organisation.

Technical architecture

The following diagram shows the first end-to-end data capture to archiving pathways to be turned on at the University of Western Sydney, covering Acquisition and Action on data (*use*) and Archiving and Advertising of data for *reuse*. Note the inclusion of a name-authority service which is used to ensure that all metadata flowing through the system is unambiguous and inked-data-ready [7]. The name Authority is populated with data about people, grants and subject codes from databases within the research services section of the university and from community-maintained ontologies. A notable omission from the architecture is integration with the Institutional Publications Repository – we hope to be able to report on progress joining up that piece of the infrastructure via a Research Hub at Open Repositories 2014.


Joined up 4A data managemnt
Institute Facilities
Free Air Co2 Experiment
WHole Tree Chambers
Eddy Flux Tower
New data!
New data!
New data!
HIEv Environmental Data Capture
IT Service Desk / Operations
Manage
Library / Research Services
Can provide basic 'R Drive' for research
Fetch data
Look up names and subject codes etc
New data?
Provide assistance in describing research outputs
Research Data Repository
Acquire & Act
Research Data Store - Archival
Mint Name Authority
Research Data Catalogue
Research Data Store - Working
Advertise
Sync
New data holdings to Advertise?
Web aggregation services
Archive
People Australia @ National Library
Research Data Australia

References

1. Burton, A. & Treloar, A. Designing for Discovery and Re-Use: the 'ANDS Data Sharing Verbs' Approach to Service Decomposition. *International Journal of Digital Curation* **4**, 44–56 (2009).

2. Androulakis, S. MyTARDIS and TARDIS: Managing the Lifecycle of Data from Generation to Publication. in *eResearch Australasia 2010* (2010).at <http://ccaeducause1.caudit.edu.au/index.php/eraust/2010/paper/view/62>
3. Sefton, P. M. The Fascinator - Desktop eResearch and Flexible Portals. (2009).at <https://smartech.gatech.edu/handle/1853/28483>
4. Kunze, J., Boyko, A., Vargas, B., Madden, L. & Littman, J. The BagIt File Packaging Format (V0.97). at <http://tools.ietf.org/html/draft-kunze-bagit-06>
5. Group, W. W. & others *RDFa Core 1.1 Recommendation*. (2012).at <http://www.w3.org/TR/rdfa-syntax/>
6. Wolski, M., Richardson, J. & Rebollo, R. Shared benefits from exposing research data. in *32 nd Annual IATUL Conference* (2011).at <http://iatul2011.bg.pw.edu.pl/proceedings/ft/Wolski_M.pdf>
7. Berners-Lee, T. *Linked data, 2006*. at <http://www.w3.org/DesignIssues/LinkedData.html>

[i] Project materials refer to the repository as a project which includes both working and archival storage as well as some computing resources, drawing a line around ‘the repository’ that is larger than would be usual for a presentation at Open Repositories.