

# Boundaryless eResearch: use the Web, use Linked Open Data

Peter Sefton<sup>1</sup>, Jim Downing<sup>2</sup>, Anna Gerber<sup>3</sup> Peter Murray-Rust<sup>4</sup>

<sup>1</sup>University of Southern Queensland, Toowoomba, Australia, peter.sefton@usq.edu.au

<sup>2</sup>University of Cambridge, Cambridge, UK, ojd20@cam.ac.uk

<sup>3</sup>University of Queensland, Brisbane, Australia, agerber@itee.uq.edu.au

<sup>4</sup>University of Cambridge, Cambridge, UK, pm286@cam.ac.uk

## INTRODUCTION

In keeping with the theme of this conference, 'no boundaries' this BoF session will explore how the World Wide Web can be used to leap the boundaries, break the barriers and storm the barricades which circumscribe and constrain eResearch. Our proposal is a simple mantra, or set of design principles for eResearch analysts and service providers: "Use the web. Use Linked Open Data". Using the web means using the World Wide Web<sup>1</sup> as it was intended, the web that Berners-Lee and collaborators created for scholarship, not merely relying on those institutional and corporate systems that wrap PDF documents and ignore or hide data. The rules, or guidelines for Linked Data<sup>2</sup> are:

- Use URIs as names for things
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
- Include links to other URIs. so that they can discover more things.<sup>2</sup>

Linked Data is a most promising approach for technically interoperable research data. The approach is not bounded by geography, an academic discipline or even by research as an activity. By using the Linked Data fabric allows the research community to benefit from tools developed in the commercial sector, and also to have their tools widely used by all sectors, in contrast with previous parochial solutions based on specialized technologies and approaches. For data to be truly interoperable, we must also hurdle the social and legal barriers presented by data licensing, and so the second part of our message is to use Open Data licenses (Science Commons Open Data Protocol<sup>3</sup> & Open Knowledge Foundation's Open Data Definition and licenses) as well, creating Linked Open Data (LOD). Without widespread open licenses, Fear Uncertainty and Doubt will prevent practical data inter-operation.

This BoF session will bring together those who are already producing and publishing Linked Open Data, and those who want to know more, to discuss LOD exemplars and challenges. It will consist of a short introductory presentation (15-20mins) on the whys and wherefores of Linked Data, followed by short presentations on examples where Linked Data has been effectively produced, and demonstrations of tools that can be used to produce and manage Linked Data. The aim will be to effectively communicate the opportunities and challenges of Linked Data through these presentations, and hopefully move towards consensus on solutions through the ensuing discussion.

## COMMUNITY CHALLENGES

Whilst the LOD principles are easy enough for a computer scientist to understand, they are more difficult for many researchers to embrace, and harder for them to implement as they do not have tools which allow them to deal with and identify data across the research life cycle. There are, broadly speaking, two main challenges:

- The web does not reach deeply enough into the scholarly process for data to have HTTP URIs. That is while there are many web-based repositories and services there is still gap between what's on a researcher's computer and the web.
- Different web systems have different ways of doing things so following the rules for linked data is not trivial, even for experts and constructing meaningful context for links is challenging.

To this end we are recommending two design patterns for eResearch systems and will discuss in this presentation how our experience with academic computing systems have led us to adopt and re-adopt these patterns:

- bringing semantically rich web systems down to the desktop and the lab so that researchers see their data linked to the web from the instant it is created, and
- using HTML for research communications from collaboration to publication, in the form of an emerging set of practices for web-based scholarship that can be applied across systems: Scholarly HTML.

One of the keys to the mantra "Use the web" is that the web is not made of monolithic systems – it is, if anything, made of links so taking this seriously does not mean attempting to build systems which are all-encompassing it means exposing as much as possible via the web. There are two key pieces of infrastructure that have recently been released by ANDS which will be a great help in this area. Register My Data<sup>4</sup> allows data collections to be described in a web-readable (ie human and machine) way and Identify My Data<sup>5</sup>, a service to help people use URIs as names for things and use HTTP URIs so that people can look up those names.

## WEB-NATIVE ERESEARCH EXAMPLES

One key line of demarcation is between the research desktop or lab computer and the Data Commons. Data are not on the web until they are put there. And while web based collaboration systems are increasingly being used and promoted the means by which researchers might collaborate and use linked data practices are not yet clear; we believe that some of the work we have done to bring the web to the desktop has some of the solutions.

## DATA NETWORKS / COLLABORATIONS

The presenters will refer to the following collaborations which we have been involved in:

**The Aus-e-Lit project<sup>6</sup>** is a NeAT-funded project that aims to address the eResearch needs of researchers involved in the study of Australian literature and Australian print culture. The project enhances and extends the existing AustLit web portal with data integration and search services, empirical reporting services, compound object authoring, editing and publishing services and collaborative annotation services.

**The OREChem project<sup>7</sup>** is a Microsoft-funded collaboration between Cambridge, Cornell, Indiana, Penn State and Southampton Universities that aims to make existing chemistry data sources available as LOD, to develop new LOD creation resources using grid computing, to develop and converge on standard ontologies for chemistry knowledge representation, and to further the state of the art in extracting semantic chemistry data from published PDF.

Talis Connected Commons & CrystalEye

**Talis Connected Commons** is an initiative whereby Talis offer free Linked Data infrastructure for open datasets. Nick Day and Jim Downing at the University of Cambridge are working to publish semantic data from the CrystalEye<sup>8</sup> system through Talis Connected Commons.

#### DESKTOP / NEAR DESKTOP SYSTEMS

Some examples of efforts to bring the semantic web closer to the desktop include repository based systems such as Islandora and the group of projects including RepoMMan, REMAP and Hydra<sup>9</sup>. These require the user to locate data and documents themselves and add them to the online repository. Closer to the desktop many content management systems use WebDAV. A notable example which is desktop but not web based is [MediaFlux](#).

Our collaborations include the following systems that implement the proposed designpattern, ICE.<sup>10</sup> and The Fascinator Desktop<sup>11</sup> both bring the web to the desktop. Lensfield – a desktop data processing environment that brings the semantic web to the process, using SPARQL and web harvesters, and allows easy publication of Linked Data through a variety of channels<sup>12</sup>. These implement the pattern of bringing the web to the eResearcher, if that means making their desktop part of the web then we aim to do so.

#### WEB SCHOLARSHIP EXAMPLES

The other main barrier to consider is that around a typical research publication. As Peter Murray-Rust puts it, “PDF is a hamburger, and we’re trying to turn it back into a cow”<sup>13</sup> and more formally in his call, with Henry Rzepa for a documents to be made available in HTML as ‘datuments’<sup>14</sup>. We will give a brief overview of work in this area.

1. Berners-lee, T. & Cailliau, R. The World-Wide Web. (1994).at <<http://eprints.kfupm.edu.sa/71757/>>
2. Berners-Lee, T. Linked Data - Design Issues. (2009).at <<http://www.w3.org/DesignIssues/LinkedData.html>>
3. Science Commons Science Commons » Protocol for Implementing Open Access Data. at <<http://sciencecommons.org/projects/publishing/open-access-data-protocol/>>
4. Grant, H. Register My Data. *Australian National Data Service* (2009).at <<http://ands.org.au/services/register-my-data.html>>
5. Grant, H. Identify My Data - Overview. *Australian National Data Service* (2009).at <<http://ands.org.au/services/identify-my-data.html>>
6. Aus-e-Lit project Aus-e-Lit: Overview. at <<http://www.itee.uq.edu.au/~eresearch/projects/aus-e-lit/>>
7. Lagoze, C. The oreChem Project: Integrating Chemistry Scholarship with the Semantic Web. *Proceedings of the WebSci'09: Society On-Line* (2009).at <<http://journal.webscience.org/112/>>
8. Crystaley project CrystalEye: Homepage. at <<http://wwmm.ch.cam.ac.uk/crystaley/>>
9. Project Hydra: Designing & Building a Reusable Framework for Multipurpose, Multifunction, Multi-institutional Repository-Powered Solutions - Georgia Tech's Institutional Repository. at <<http://smartech.gatech.edu/dspace/handle/1853/28496>>
10. Sefton, P. The Integrated Content Environment for Research and Scholarship. *ICE Website* (2006).at <[http://ice.usq.edu.au/introduction/ice\\_rs.htm](http://ice.usq.edu.au/introduction/ice_rs.htm)>
11. Sefton, P. & Lucido, O. The Fascinator: a lightweight, modular contribution to the Fedora-commons world. (2009).
12. Downing, J. lensfield - Google Code. *Project website* at <<http://code.google.com/p/lensfield/>>
13. Rusbridge, C. Science publishing, workflow, PDF and Text Mining. *Digital Curation Blog* (2008).at <<http://digitalcuration.blogspot.com/2008/05/science-publishing-workflow-pdf-and.html>>
14. Murray-Rust, P. & Rzepa, H.S. The Next Big Thing: From Hypermedia to Datuments. *Journal of Digital Information* 5, 248(2004).