RESEARCH ARTICLE

Pathways, parallels and pitfalls: the Scholarly Web, the ESRC and Linked Open Data

Antonina Lewis and Peter Neish

eScholarship Research Centre, The University of Melbourne, Melbourne, Australia

CONTACT Peter Neish peter.neish@unimelb.edu.au

Antonina Lewis ORCID 0000-0001-9540-5438
Peter Neish ORCID 0000-0002-7171-2334

Abstract

This paper highlights key principles that the eScholarship Research Centre (ESRC) shares with the Linked Open Data (LOD) community, primarily: relationship-centric contextualisation of information resources; and a commitment to producing and publishing sustainable, standards-based data outputs suitable for machine-based interchange. This paper illustrates how these principles have enabled the ESRC, without pursuing LOD as a specific end, to hold a path close to the Linked Data road. We also note that, by extension, this has positioned the ESRC ready to translate many of its resources into Linked Open Data formats in exchanges where complementary technologies and ontology mappings are available.

Keywords
Linked open data; metadata; semantic web; scholarly web; sustainability; schema.

Introduction

The eScholarship Research Centre (ESRC) has a strong belief in the potential of the Scholarly Web: an information ecology – albeit as yet unrealised (Van de Sompel & Nelson, 2015) – that is truly supportive of highly distributed activities of scholarship and scholarly communication as they occur in today's digital and online environments. One among many organisations journeying along different paths and at varying speed toward the collaborative ideal of the Scholarly Web, the ERSC has long been an advocate for the benefits of creating and publishing relationship-rich structured data expressed in standards-based formats.

To date, the theory and groundwork for data exchange, rather than the actual mechanics of interoperability, have been the primary focus for the ERSC's work in this area: as an 'out of the box' service the Centre does not produce Linked Open Data under the strictest W3C definition of the term. In part, this is a legacy consequence of an early technical decision not to utilise the RDF format in the ESRC's data management tools. Instead, the bulk of the Centre's published data resources are generated by way of relational databases which are used to produce machine readable documents in HTML and in a number of XML formats. While not producing Linked Open Data by default, this approach has facilitated collaborations with other agents that have seen a number of ESRC datasets made available as Linked Data via third party applications such as HuNI, the Humanities Networked Infrastructure (https://huni.net.au/#/partners). More recently, the ESRC has begun utilising

JSON as a data format to facilitate development of discovery tools for the datasets it manages, with the potential to bring its own data publications closer to alignment with 'true' Linked Open Data outputs.

Context

The ESRC has its origins in archival rather than library practice, with collections of personal papers and organisational archives strongly featured in the Centre's catalogue of research projects and contract consultancies. The historical predecessors of the Centre (its first incarnation being the Australian Science Archives Project), as well as the professional specialisations of many staff over the length of this history, have held an archives perspective and have seen the ESRC maintain a continuing regard for 'relationships' (historic, active, or conceptual) as a core building block for making information meaningful, with information understood to embrace both primary resources and their metadata.

In common with a traditional library environment, the Centre seeks to share and connect information resources with users. More broadly, the Centre involves itself with research questions relating to social informatics and the sustainability of knowledge, particularly in relation to digital and online environments. The work of the ESRC recognises – and often struggles with – the effects of online information instability and decay. In the online context, information instability generally results from a conscious act on the part of resource maintainers (for example, a government department removing pages created by its predecessor in the wake of electoral or other administrative change), whereas information decay is the result of neglect rather than active intervention. The ripple effects and repercussions of online phenomena such as content drift and link rot are pressing issues for the scholarly citation of web resources and data, but they also have more immediate social and legal impacts (such as the procedural change introduced by the US Supreme Court website, as summarised at http://freegovinfo.info/node/10449). Building sustainability into the data underpinning the web necessarily means recognising change while being flexible enough to accommodate it.

The institutional reporting line for the ESRC is through the Research and Collections arm of the University of Melbourne's library services. Predictably, the grouping of disparate University collections under one banner, and parallel institutional ambitions for creating aggregate repositories of content, have posed some hurdles. Notably, issues have arisen with creating meaningful links between collections and in maintaining the integrity and consistency of source metadata. For example, in the Digitised Collections repository (https://digitised-collections.unimelb.edu.au), collection identifiers for items from the University Archives were initially being captured in different places (sometimes as part of the *dc.title*; sometimes as a *dc.description* element), and in different ways (sometimes at item level, other times only a series identifier). Capturing key identifying metadata in such a non-standardised manner presents issues for researchers wishing to cite or pursue further research, particularly if no link to the archival catalogue (where associated context and collections might be found) is included.

These problems are not unique to the University of Melbourne experience, stemming in part from dissimilarity in the ways that different functional areas and disciplines across the spectrum of libraries, archives and cultural collections choose to prioritise and express metadata for collection materials. However, despite the seemingly inevitable professional

differences of opinion, standards, and schema, many commonalities are shared. Both libraries and archives are now operating firmly in the realm of information services, with core business in the practice of describing, connecting, safeguarding, and sharing knowledge resources. Similarly, the challenges (both positive and negative) arising from public expectations for online and on-demand delivery of these resources are a point of commonality for the two professions.

If the first step in providing usable information to an audience is making it discoverable, the second step is making it accessible. Openness is a common concern for library and archive workers (as is its flipside, surveillance). In large part – although not exclusively – a point of difference for information services workers choosing to work in public libraries and archives is the self-identified commitment to public good and equity of access to information. The 2008 report *Enriching communities: The value of public libraries in New South Wales* observes: 'Social wellbeing was … strongly linked to public library collections, which were seen [by surveyed library staff] to: a) Address disadvantage by ensuring free and equitable access to collections for all community members' (Library Council of New South Wales, 2008, p. 9). This commitment to connected, open, and accessible public knowledge resources informs, if not drives, the work of many contemporary libraries and archives. Accordingly, trust is critical to the ability to perform our roles most effectively whether as safe spaces or as repositories of reliable evidence; the perception of the library (or archive) as trustworthy by its user community is essential. Without that relationship in place, no matter how high the quality of information on offer, it will not be used to its full potential.

Van de Sompel and Nelson (2015) accurately note that, despite bigger intentions, it is still the case that 'Most scholarly nodes can best be characterized as stand-alone portals, destinations on the web, rather than infrastructural buildings blocks in a global, networked scholarly communication system,' and we acknowledge that at the present time the online resources published by the ESRC are relatively siloed insofar as they are not capable of unmediated cross-exchange. The ESRC is able to provide information relay from its datasets (with data from a number of resources being harvested by the National Library of Australia's Trove, for example). However, such transfer is largely uni-directional: our online resources do not yet have mechanisms that would situate their own scholarly citations outside the closed system of the underlying databases. Attending to disambiguation, broken links, and the identification and correction of citation errors are processes that still require a high degree of human intervention. As such, there is a high maintenance cost incurred in sustaining the currency of the data relating to external resources that are cited within them.

The promise of linked data as a mechanism to open up global networks of information by improving the visibility and built-in redundancy of data resources is a long way from being fully realised. In the interim, the logic and reality of Linked Data as a part of the networked information landscape is validated by the efforts of library, archive and cultural sectors worldwide in bringing their data to the table. For example: the Bibliographic Framework (BIBFRAME) initiative (https://www.loc.gov/bibframe/) seeks to use Linked Data principles to redefine the description of bibliographic records; the Linked Archival Metadata project and Architypes community group (https://www.w3.org/community/architypes/) are exploring ways to better encode information about archives using linked data; and many cultural collections are being aggregated and exposed using Linked Data (such as Europeana and the Digital Public Library of America). These examples are diverse and exciting in their willingness to radically rethink descriptive needs in light of changing technologies. The use

of Linked Data has the potential to achieve frictionless research communication and collaboration by linking and combining data unambiguously across collections.

Linked Data

Linked Data is a natural extension of the World Wide Web, except that instead of linking HTML documents together, data are linked. Linked Data consists of a set of best practices for publishing and connecting structured data on the Web (Bizer, Heath, & Berners-Lee, 2009). Technically, Linked Data refers to data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and it can in turn be linked to from external data sets. By creating the links in a standard way, machines are able to read the relationships between data objects and traverse the web of information.

The term Linked Data was first coined by Tim Berners-Lee in his Linked Data design note (Berners-Lee, 2006) in which he outlines four rules for publishing data on the web:

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

These have become known as the 'Linked Data principles' and form the underpinnings of the Semantic Web. There have been differing interpretations as to what exactly a Semantic Web would look like, but according to Berners-Lee (2000) 'The first step is putting data on the Web in a form that machines can naturally understand, or converting it to that form. This creates what I call a Semantic Web – a web of data that can be processed directly or indirectly by machines'.

Berners-Lee's design note was subsequently updated in 2009 to specifically include the concept of Linked Open Data, that is, Linked Data that is published with an open licence, which does not impede its reuse for free. Berners-Lee also included the concept of a star rating for data (Figure 1) to indicate that there is a continuum of Linked Data.

★	Available on the web (whatever format) *but with an open licence, to be Open Data*
★★	Available as machine-readable structured data (e.g. excel instead of image scan of a table)
★★★	as (2) plus non-proprietary format (e.g. CSV instead of excel)

★★★★	All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
★★★★★	ll the above, plus: Link your data to other people's data to provide context

Figure 1. Linked Open Data star rating (http://www.w3.org/DesignIssues/LinkedData.html).

The Linked Open Data Cloud (http://lod-cloud.net/) provides a dramatic visualisation of the number of Linked Open Data sources there are on the web. There is no doubt that the number of data sources has grown rapidly since 2007 (Figure 2).

Impressive as this growth is, the number of datasets does not tell the entire story. The quality of the data and the quality of the links are important factors. In the latest Linked Open Data cloud, only fifty-six percent of datasets link to another dataset, while the remaining are only targets for links (Schmachtenberg, Bizer, & Paulheim, 2014). This lack of connectivity between datasets limits the possibilities to automatically traverse the information, as nearly half of all datasets are effectively dead ends.

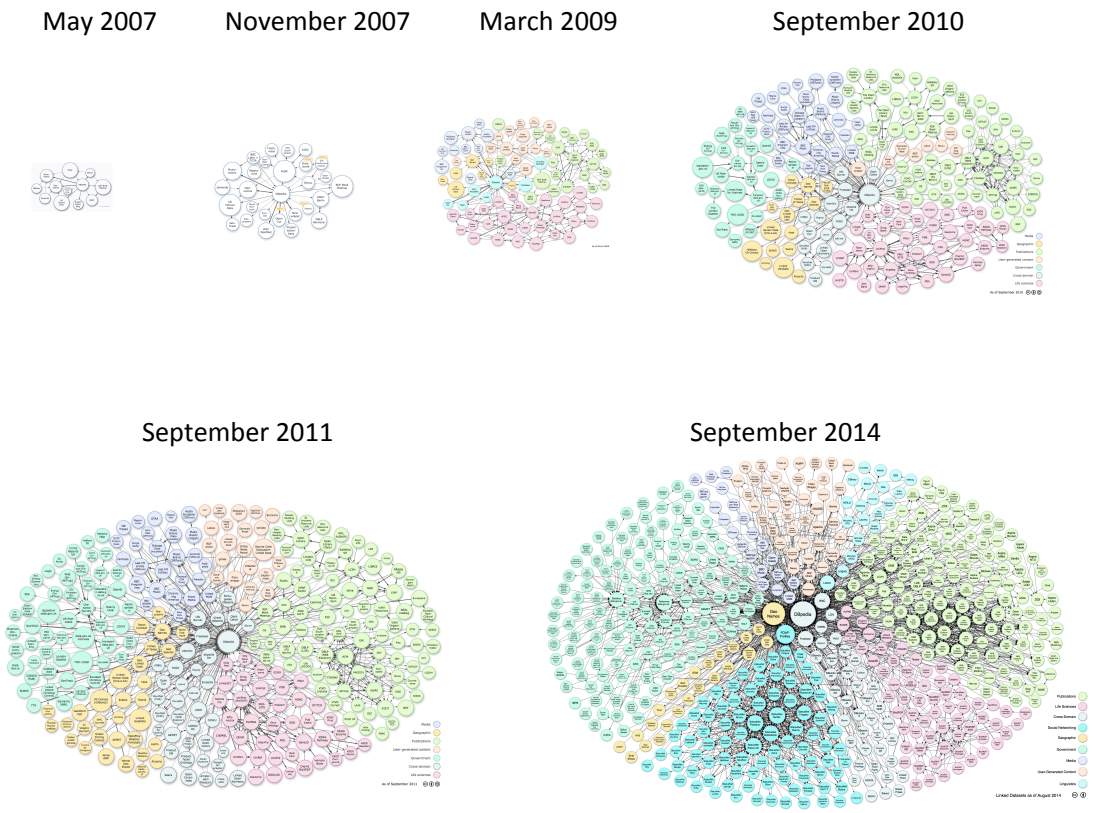

Figure 2. Growth of the Linked Open Data Cloud (http://lod-cloud.net/).

Another consideration is the extent to which data sets share a widely used vocabulary. If different datasets use different vocabularies then it is much harder or impossible to infer that objects are the same without additional cross-walking or mapping of the values. Of all the vocabularies encountered in the cloud, less than half are used by more than one dataset (Schmachtenberg et al 2014).

Despite the growth in Linked Data, genuine examples of Web-wide Linked Data integration are hard to find. Rather than a frictionless, interconnected web of data, we are more likely to find linked data being used in specific applications and domains where the benefits are realised by a single organisation or community (Neish 2015).

ESRC

A more detailed history of the ESRC is provided elsewhere in this issue. Of interest in this article are the parallels between the history of the ESRC and the evolving World Wide Web and Linked Open Data movement (
Figure 3).

The World Wide Web was released to the public in August 1991, which was followed by a period of rapid escalation of web technologies, including web browsers and underlying standards. The ESRC and previous incarnations have been early adopters of web technologies. The Australian Science Archives Project embraced the web as a means of disseminating information and 1994 saw the release of the Bright Sparcs website, available continuously for over 20 years (now incorporated into the Encyclopaedia of Australian Science, http://www.eoas.info/background.html). The use of the web by the ESRC has always been based on the fundamental principles of preservation and discoverability. To this end, the ESRC systems have made use of persistent URLs for web resources and using unique identifiers when linking to external data. Content has been well structured according to current web standards and the data have been kept separate from the formatting and presentation, enabling new outputs to be developed relatively easily in a rapidly changing World Wide Web (Figure 3).

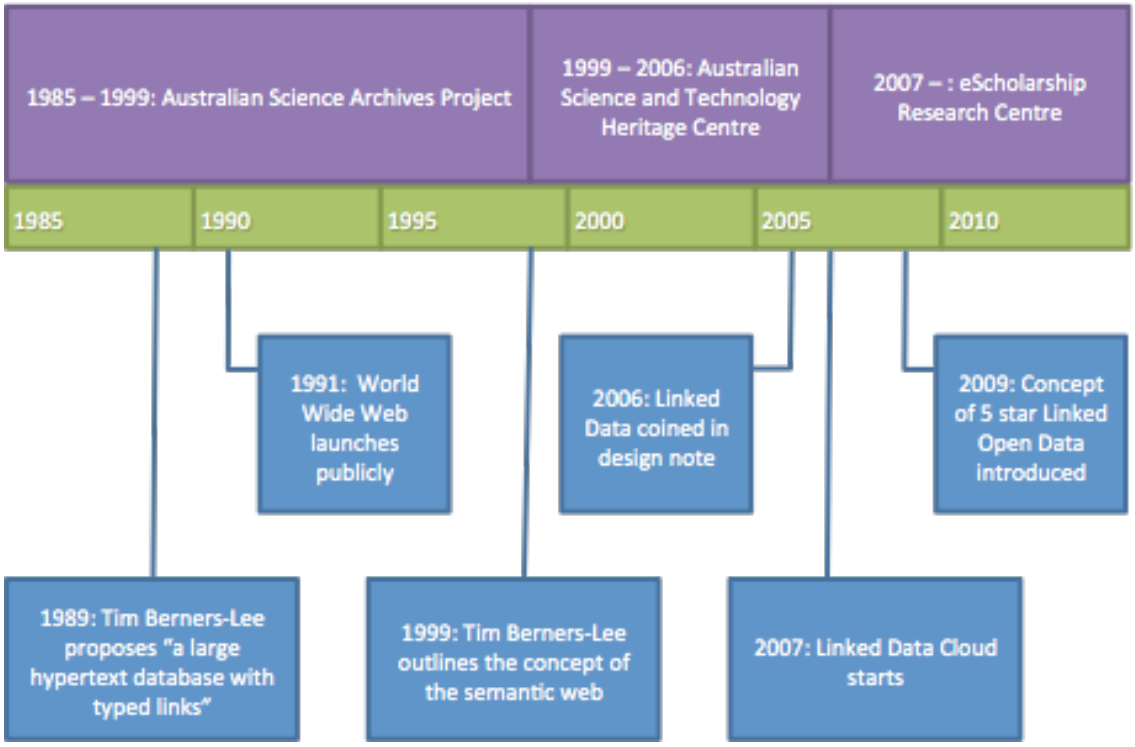

Figure 3. Timeline of the ESRC and related developments in the World Wide Web and Linked Data.

Despite being an early adopter of web technology, the ESRC chose not to use RDF as an underlying data format. While the benefits to disseminating content via the web were immediately obvious (and could be measured in hits and search rankings), the benefits of adopting Linked Data were not as clear. Despite being hyped as the next big thing, there are still relatively few concrete examples that demonstrate the benefits of using Linked Data. Another barrier has been the slow development the standards and idioms for publishing Linked Data. For example, the mechanism for dereferencing URIs (a key part of Linked Data) was not decided until 2005 and is still a lively subject of debate (W3C Technical Architecture Group, 2007).

Rather than utilising RDF, datasets compiled or otherwise curated by the ESRC – while robustly modelled, highly structured and in many ways adhering to underlying principles of Linked Data – have been stored in relational databases: the Online Heritage Resource Manager (OHRM) and the Heritage Documentation Management System (HDMS). Outputs of these databases are predominantly expressed either as HTML, enabling push-button generation of publication-ready human readable web pages; or as XML, providing machine-readable data for query or exchange. The XML renditions are constructed utilising a number of recognised schemas, with the schema depending on the type of parent record. Archival records held in the HDMS database will be represented as Encoded Archival Description (EAD); context and provenance entities in the OHRM as an expanded ESRC-variant of Encoded Archival Context for Corporate Bodies, Persons and Families (EAC-CPF) (ESRC-EAC); and linked publications or bibliographic records as Metadata Object Description Schema (MODS).

The genesis and ongoing evolution of the OHRM and HDMS are discussed in more depth elsewhere in this issue. From very early on, the OHRM and HDMS databases were designed to be used for the generation of HTML outputs, enabling a simple dissemination path for curated data to be made publicly available as web resources. However, the Centre increasingly recognised that these outputs were 'not readily sharable with other organisations or other resources operating in a similar sphere, resulting in limited reuse of this data and increased potential for duplication of work' (Smith & La Rosa, 2015). This realisation has led the ESRC to a greater commitment to providing database exports in alternate schema. Using the MODS, EAD, and EAC-CPF schema as a basis for XML outputs for these resources (in parallel with the longstanding web presentation HTML outputs) equips not only the ESRC, but also its formal collaborators – and, in theory, any interested external party – with consistently shared logic from which to make better use of the data underpinning the published resources.

Our goal, like that of the Scholarly Web and LOD communities, is that the projects we contribute to will come to include not only those applications and services used to display and navigate individual datasets published by the ESRC, but also the capacity for others to build services that run across datasets and which are interoperable with collections beyond the Centre. While we cannot yet fulfil all the dreams of the Scholarly Web, we can (and do) put measures in place that will facilitate the use of data beyond individual research projects. In the case of the ESRC, we achieve this by providing data that are persistent (with stable URLs, and more recently incorporating DOIs in some projects), data that are contextualised (relationship rich), and data that has been pre-formulated, ready to be shared via markup languages that wrap and describe that data and its attributes in a consistent fashion.

Conclusion

As the ESRC does not compile its data resources in graph stores or represent datasets in RDF as a default, it had for many years remained in orbit around the third star of the Berners Lee star rating for Linked Data (Figure 1). Since January 2014, with the adoption of JSON-LD as a W3C standard, the constellations have shifted and the ESRC has found itself aligned with the elusive fourth star. While the ESRC was not actively pursuing the Linked Data path and was technically non-conformant to the W3C ideal, the principles underpinning both Linked Data and its Open counterpart (LOD) have long been at the heart of information design at the Centre.

Linked Open Data and the Scholarly Web are long-term propositions and the daily reality for the majority of library, archive, and research bodies (the ESRC included) is not an exercise in finding perfection. As libraries and archives well know, the balance between 'information' and 'services' is not easy to achieve, and the question of how to find this balance and perform our professional functions effectively, sustainably, cooperatively, and at scale is something that engages professionals in both arenas. Our path is an ongoing and iterative process of making information resources available in the best way we can, while ensuring that the underlying data is both robust and flexible: suitable to be repurposed when a better way arises; as it inevitably will.

References

Berners-Lee, T. (2006). *Linked Data*. Retrieved from https://www.w3.org/DesignIssues/LinkedData.html

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data – the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, 205-227.

Library Council of New South Wales. (2008). *Enriching communities: the value of public libraries in New South Wales*. Sydney, NSW: Author.

Neish, P. (2015). Linked Data: What is it and why should you care? *Australian Library Journal 64,* 3-10. doi:10.1080/00049670.2014.974004.

Schmachtenberg, M., Bizer, C., & Paulheim, H. (2014). *State of the LOD Cloud 2014.* Retrieved from http://linkeddatacatalog. dws. informatik. unimannheim. de/state.

Smith, A., & La Rosa, Marco. (2015, July). *Why standards are critical: Graphing knowledge by building on well established entity relationship standards can look well into the future*. Paper presented at DH2015 panel session *The History of Science in the Age of Networked Digital Humanities*. Retrieved from http://dh2015.org/abstracts/

Van de Sompel, H., & Nelson, M. L. (2015). Reminiscing about 15 years of interoperability efforts. *D-Lib Magazine 21* (11/12*).* doi:10.1045/november2015-vandesompel.

W3C Technical Architecture Group (2007) *Dereferencing HTTP URIs.* Retrieved from https://www.w3.org/2001/tag/doc/httpRange-14/2007-05-31/HttpRange-14

Notes on contributors

Antonina Lewis is a professional archivist. She works at the University of Melbourne's eScholarship Research Centre as a Research Archivist and as Program Manager for the Find and Connect Web Resource. Antonina holds a BA in Creative Arts (Griffith University) and a PhD in Communication and Creative Arts (Deakin University). Her research interests include: dynamics of archives and power; the renegade; and the roles of creative and political actors in constructing socio-cultural narratives.

Peter Neish is the Research Data Curator at the University of Melbourne where he works in partnership with researchers on a wide range of data management projects. He has previously worked at the Victorian Parliamentary Library and the Royal Botanic Gardens Melbourne, using his background as a researcher and computer scientist to make databases and information more available, standards-based and linked. He has contributed to national and international biodiversity initiatives and data transfer standards.

Page 6: [1] Deleted	Peter Neish	15/12/2016 10:04:00

Figure 3

University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Lewis, A; Neish, P

Title:
Pathways, parallels and pitfalls: the Scholarly Web, the ESRC and Linked Open Data

Date:
2016-08-01

Citation:
Lewis, A. & Neish, P. (2016). Pathways, parallels and pitfalls: the Scholarly Web, the ESRC and Linked Open Data. AUSTRALIAN LIBRARY JOURNAL, 65 (3), pp.224-231. https://doi.org/10.1080/00049670.2016.1209075.

Persistent Link:
http://hdl.handle.net/11343/122798

File Description:
Accepted version