Funding Proposal

Expressing research output through linked data

Proposed Research

Understanding the impact of any research project is crucial, but is also a perennial challenge for both funders and researchers. Statistics based models used to generate figures that quantify impact are usually based upon data that is ambiguous, lacks transparency, and is inconsistent across different disciplines. Here we propose an implementation of a linked data ontology designed for describing research output. We hypothesize that using linked data as a vehicle for detailing research output will take steps toward addressing the issues identified with current methods for quantifying impact statistically, and will lay the foundations for further relevant research.

Context

The impact of research, researchers, publications, and institutions is often calculated algorithmically using a number of metrics. These are usually based upon bibliographic techniques applied to research repositories and citation indexes. Specific techniques for calculating impact include:

- Citation counts
- Impact factors
- H-indexes

Citation counts count the number of citations to a particular piece of published scholarly literature from other published works. *Impact factors* are a calculated value that represent the average number of citations to an article published in any given journal (an inference being that differing journals have differing levels of 'impact' based upon the average number of citations). *H-indexes* are calculated by measuring the volume of published papers within a given time period, in a ratio against the number of citations attributed to the same works. Each of these measures are usually calculated based on data held by a number of repositories and indexes, examples include Google Scholar, Thompson Reuters' Web of Knowledge, and Scopus.

These methods for viewing the impact of research, or researchers have flaws and can misrepresent impact, or represent it ambiguously. For instance, citation counts don't give context to citations (a paper may be cited many times because it is an example of *how not to do it*) (Garfield, 1979). Also a citation count can only count the citations in publications that the given index is aware of (nobody indexes *everything*). H-indexes provide a valuable measure, but in some circumstances may not represent the true 'value' of a given researcher's work (Bornmann and Hans-Dieter, 2007). Also H-indexes vary drastically between different sources. (Bar-Ilan, 2007).

Beginning in the 1950s, Eugene Garfield pioneered citation indexes with this work at the Institute for Scientific Information (ISI), which in turn facilitated effective use of the bibliographic techniques mentioned above. Garfield in fact also invented the concept of impact factors. Although these bibliographic techniques are extremely useful, and in comparison with the entirely manual systems originally devised by the ISI are incredibly sophisticated, none of the existing techniques when viewed in isolation can reliably inform of any particular research's impact holistically. Garfield himself (1998) has warned of the dangers around using impact factors saying *"The source of much anxiety about Journal Impact Factors comes from their misuse in evaluating individuals"*.

Understanding impact is of particular importance in order to make strategic investment in research. Similarly, pathways to impact are crucial. One relatively recent development is the shift toward more open access

journals. Research in open access publications usually has a higher impact rating than closed access journals (Harnad and Brody, 2004). The UK government has in the last 12 months elected to make sure that publicly funded research is published in open access journals, as per the recommendations made by Dame Janet Finch in her 2012 report commissioned by the UK government *Expanding Access to Published Research.*

Although most statistics-based measures of impact are based upon data held by 3rd party repositories (Google Scholar, etc) the majority UK research institutions usually catalogue their research independently as well. These internal databases are usually used for internal governance purposes and to support initiatives such as Research Excellence Framework submissions. Several UK universities (including Lancaster, Bristol, St Andrews, Edinburgh, Dundee, York) use a common platform for storing this data called Pure. Although each implementation of Pure holds a wealth of information, it isn't generally accessible.

Linked data is a technology, it is a system for implementing the 'web of data' (as opposed to a web of documents). The key distinction between data and documents is that data can be understood by computers, which has profound implications. Linked data utilises *ontologies* to describe things on the web (this may be a person, place, object, *type* of object, or anything else). These ontologies may be applied to any abstract entity, and are the data is exposed using technologies such as RDF or XML. Many different ontologies, for describing many different things, have been implemented on the web already. One such ontology is the BIBO ontology, this is at the core of many different projects, including the Digital Object Identifiers or DOIs (the most prevalent standard for referencing academic literature in digital realms).

We propose a research project to design and build a new linked data ontology specifically for documenting research output.

Bar-Ilan, J., 2007. Which h-index? — A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74(2), pp.257–271. Available at: http://www.springerlink.com/index/10.1007/s11192-008-0216-y [Accessed February 27, 2013].

Bornmann, L. & Daniel, H., 2007. What Do We Know About the h Index ? *Journal of the American Society for Information Science and Technology*, 58(9), pp.1381–1385.

Garfield, E., 1979. Is citation analysis a legitimate evaluation tool? *Scientometrics*, 1(4), pp.359–375. Available at: http://onlinelibrary.wiley.com/doi/10.1002/cbdv.200490137/abstract [Accessed March 2, 2013].

Garfield, E., 1998. The impact factor and using it correctly. *Der Unfallchirurg*, 48(2), pp.413–414. Available at: http://www.garfield.library.upenn.edu/papers/derunfallchirurg_v101(6)p413y1998.pdf [Accessed March 2, 2013].

Harnad, Steven; Brody, T., 2004. Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. *D-Lib Magazine*, 10(6). Available at: http://www.dlib.org/dlib/june04/harnad/06harnad.html.

Objective

To develop an open and extensible linked data ontology designed to allow universities to describe their research output independently of citation indexes or literature repositories. We believe this will underpin a new way to understand research impact including taking steps to address the flawed impact measurement systems described above. The ontology alone is akin to a schematic and isn't immediately useful. To be have any utility it requires: (1) data to be exposed through it and; (2) applications to be built on top of it. We propose using the data hosted in the Pure implementation at Lancaster University as test data for the ontology; a middleware software component will be developed to facilitate this. As a proof of concept we will also develop a visualisation

system to showcase the ontology, we anticipate future research and development will explore other uses of the ontology. This is fundamental research, providing the foundation for further research.

Methodology

We propose a 'design/build/demo' approach to this research with three components:

1. Linked data ontology
2. Middleware component
3. Proof of concept visualisation application

As the core of the research, the ontology will demand the majority of attention. A period of desk-based research uncovering the intricacies of different approaches to compiling/using citation indexes will inform the specification/development of the linked data ontology. After this initial exploratory period and with a draft specification in place, the middleware component will be developed, allowing the data that already resides in Pure to ‘plug in’ to the ontology. Finally the visualisation system will be built in order to showcase the potential uses for the ontology.

Beyond the ‘design/build/demo’ paradigm we will hold a series of workshops and interviews at the conclusion of the project in order to collect a breadth of opinion from researchers, developers, and university management. These workshops will serve to explain, using the visualisation application, the potential for further exploitation of the new ontology. Interviews will be conducted in order to understand how workshop participants see the linked data ontology contributing to their work, and what potential applications there are for utilising the ontology further. The workshops will be hosted by a professional facilitation team. Interviews will be conducted to collect data from the workshop participants. Qualitative analysis/coding techniques will be used to interpret the data.

Programme of Work

We propose research taking place over a 26 week period, details of the work programme are shown below.

Week 1-8	Desk research, producing a detailed report on citation indexes, repositories, and statistics based impact measures
Week 9-16	Specification/development of linked data ontology
Week 17	Development of middleware component
Week 18-22	Development of visualisation application
Week 23	Workshops and interviews
Week 24-26	Analysis of workshops, forming conclusions, producing outcome report

Importance

The potential applications for linked data are far reaching and have profound consequences, arguably linked

data could be as important for the web, as the web was for the internet (Bizer, Heath & Berners-Lee, 2009; TED, 2009). This research explores one application of linked data. The importance of this research is derived from the importance of research itself. 'Good' research is valuable in any context, however being able to quantify *why* and *how* it is 'good' is a huge challenge. In order to mitigate the difficulty of objectively quantifying how 'good' a particular piece of research is, impact factors are a valuable tool. Our research intends to augment and extend existing paradigms for understanding research impact, by doing so we believe richer understandings will emerge. By understanding impact in new and more informative ways, funding bodies can operate more strategically. Furthermore researchers, institutions, and other interested parties will have access to the same data which may be used in a plethora of ways; the workshops that conclude the project are in part a sense-making exercise, to explore these potential applications of the new ontology.

Bizer, C., Heath, T. & Berners-Lee, T., 2009. Linked data-the story so far. *International Journal on Semantic …*. Available at: http://www.igi-global.com/article/linked-data-story-far/37496 [Accessed March 4, 2013].

TED, 2009. Tim Berners-Lee on the next Web. Available at: http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html [Accessed March 2, 2013].

Academic Impact

Although literature around measuring citation indexing, measuring research impact, and linked data is comprehensive, there is currently no research exploring how the three fields of study may influence or interact with each other. As part of this research we will develop a proof of concept application to explore how linked data may be used as a tool generating richer understanding of impact. The linked data ontology itself however is the major carrier of academic impact, and will provide the foundation of future research.

Non-Academic Impact

We believe that this project will provide the foundations for future research that will have profound implications for how research is communicated and made relevant, particularly to non-academics, and in non-academic contexts.

Impact Summary

Traditional models for indexing scholarly literature and research are steeped in a rich history, and follow traditions established over decades and centuries. The foundational work into citation indexes, beginning with Eugene Garfield's work, is one part of this. The volume and breadth of academic journals and conferences is another integral part of this landscape. These traditions are incredibly important and serve the research community very well. However, the general approaches to how research is understood, perceived, communicated, and analysed, are not in keeping with the technological or societal zeitgeist. Crowdsourcing, wiki principles, open data principles, open source principles are but some of the ideals that struggle to come to the fore as part of established practices. Furthermore, knowledge transfer from academic institutions to those who research is relevant to, is a continuing and challenging problem (Stevens and Bagby, 2001).

As well as improved understanding of impact, which is directly relevant to research institutions, we believe that this research into how linked data can be used to express research output will provide the technological

foundation to allow the research community to share its knowledge better, by using some of the principles mentioned above. The implications of this shift towards openness and collaboration include stimulating serendipitous connections (Makri et al., 2011), facilitating the un-designed creation of new communities of practice (Sherer, Shea & Kristensen, 2003; Wenger, 2000), and creating a more fertile environment for interdisciplinary innovation (Blackwell et al., 2010).

Blackwell, A. et al., 2010. *Creating value across boundaries*, Available at: http://www.nesta.org.uk/library/documents/creating_value_across_boundaries_may10.pdf.

Makri, S. et al., 2011. Proceedings of the 1st International Workshop on Encouraging Serendipity in Interactive Systems. Available at: http://discovery.ucl.ac.uk/1339629/ [Accessed March 1, 2013].

Sherer, P., Shea, T. & Kristensen, E., 2003. Online communities of practice: A catalyst for faculty development. *Innovative Higher Education*, 27(3), pp.183–195. Available at: http://www.springerlink.com/index/L594153X33H75075.pdf [Accessed March 1, 2013].

Stevens, J.M. & Bagby, J.W., 2001. Knowledge Transfer from Universities to Business: Returns for all Stakeholders? *Organization*, 8(2), pp.259–268. Available at: http://org.sagepub.com/cgi/doi/10.1177/1350508401082012 [Accessed March 1, 2013].

Wenger, E., 2000. Communities of Practice and social learning systems. *Organization*, 7(2), pp.225–246. Available at: http://org.sagepub.com/content/7/2/225.short [Accessed March 4, 2013].

Pathways to Impact

The workshops that conclude the programme of work for this project are key to impact pathways. The purpose of the workshops is primarily to collect data from a broad range of sources, around potential applications of the linked data ontology. This will serve to design future research projects based upon the data collected in the workshops.

Due to the foundational nature of this research, it has very limited *direct* non-academic impact. Therefore an important task, which the workshops will also assist in, is communicating the potential for this technology to other researchers.

We believe the best way of forging pathways to impact for this research is by effectively communicating the potential of this technology, and uncovering unanticipated applications of it.