Future Generation Computer Systems 29 (2013) 599–611

Contents lists available at SciVerse ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Why linked data is not enough for scientists

Sean Bechhofer [a,*], Iain Buchan [b], David De Roure [d,c], Paolo Missier [a], John Ainsworth [b], Jiten Bhagat [a], Philip Couch [b], Don Cruickshank [c], Mark Delderfield [b], Ian Dunlop [a], Matthew Gamble [a], Danius Michaelides [c], Stuart Owen [a], David Newman [c], Shoaib Sufi [a], Carole Goble [a]

[a] School of Computer Science, University of Manchester, UK
[b] School of Community Based Medicine, University of Manchester, UK
[c] School of Electronics and Computer Science, University of Southampton, UK
[d] Oxford e-Research Centre, University of Oxford, UK

ARTICLE INFO

Article history:
Received 8 March 2011
Received in revised form
18 July 2011
Accepted 5 August 2011
Available online 19 August 2011

Keywords:
Research object
Linked data
Reproducibility
Reuse
Sharing
Publishing

ABSTRACT

Scientific data represents a significant portion of the linked open data cloud and scientists stand to benefit from the data fusion capability this will afford. Publishing linked data into the cloud, however, does not ensure the required reusability. Publishing has requirements of provenance, quality, credit, attribution and methods to provide the *reproducibility* that enables validation of results. In this paper we make the case for a scientific data publication model on top of linked data and introduce the notion of *Research Objects* as first class citizens for sharing and publishing.

1. Introduction

Changes are occurring in the ways in which research is conducted. Within wholly digital environments, methods such as scientific workflows, research protocols, standard operating procedures and algorithms for analysis or simulation are used to manipulate and produce data. Experimental or observational data and scientific models are typically "born digital" with no physical counterpart. This move to digital content is driving a sea change in scientific publication, and challenging traditional scholarly publication. Shifts in dissemination mechanisms are thus leading towards increasing use of electronic publication methods. Traditional paper publications are, in the main linear and human (rather than machine) readable. A simple move from paper-based to electronic publication, however, does not necessarily make a scientific output decomposable. Nor does it guarantee that outputs, results or methods are reusable.

Current scientific knowledge management serves society poorly, where for example the time to get new knowledge into practice can be more than a decade. In medicine, the information used to support clinical decisions is not dynamically linked to the cumulative knowledge of best practice from research and audit. More than half of the effects of medications cannot be predicted from scientific literature because trials usually exclude women of childbearing age, people with other diseases or those on other medications. Many clinicians audit the outcomes of their treatments using research methods. This work could help bridge the knowledge gap between clinical trials and real-world outcomes if it is made reusable in wider research [1].

As a further example from the medical field, there are multiple studies relating sleep patterns to work performance. Each study has a slightly different design, and there is disagreement in reviews as to whether or not the overall message separates out cause from effect. Ideally the study-data, context information, and modelling methods would be extracted from each paper and put together in a larger model – not just a review of summary data. To do this well is intellectually harder than running a primary study – one that measures things directly. This need for broad-ranging "meta-science" and not just deep "mega-science" is shared by many domains of research, not just medicine.

Studies continue to show that research in all fields is increasingly collaborative [2]. Most scientific and engineering domains would benefit from being able to "borrow strength" from the outputs of other research, not only in information to reason

* Corresponding author. Tel.: +44 161 275 6282; fax: +44 161 275 6236.
E-mail address: sean.bechhofer@manchester.ac.uk (S. Bechhofer).

0167-739X/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2011.08.004

over but also in data to incorporate in the modelling task at hand. We thus see a need for a framework that facilitates the reuse and exchange of digital knowledge. Linked Data [3] provides a compelling approach to dissemination of scientific data for reuse. However, simply publishing data out of context would fail to: (1) reflect the research methodology; and (2) respect the rights and reputation of the researcher. Scientific practice is based on publication of results being associated with provenance to aid interpretation and trust, and description of methods to support reproducibility.

In this paper, we discuss the notion of Research Objects (ROs), semantically rich aggregations of (potentially distributed) resources that provide a layer of structure on top of information delivered as Linked Data. An RO provides a container for a principled aggregation of resources, produced and consumed by common services and shareable within and across organisational boundaries. An RO bundles together essential information relating to experiments and investigations. This includes not only the data used, and methods employed to produce and analyse that data, but also the people involved in the investigation. In the following sections, we look at the motivation for linking up science, consider scientific practice and look to three examples to inform our discussion. Based on this, we identify principles of ROs and map this to a set of features. We discuss the implementation of ROs in the emerging Object Reuse and Exchange (ORE) representation and conclude with a discussion of the insights from this exercise and critical reflection on Linked Data and ORE.

2. Reproducible research, linking data and the publication process

Our work here is situated in the context of *e-Laboratories*, environments that provide distributed and collaborative spaces for e-Science, enabling the planning and execution of in silico and hybrid studies—processes that combine data with computational activities to yield research results. This includes the notion of an e-Laboratory as a traditional laboratory with on-line equipment or a Laboratory Information Management System, but goes well beyond this notion to scholars in any setting reasoning through distributed digital resources as their laboratory.

2.1. Reproducible research

Mesirov [4] describes the notion of Accessible Reproducible Research, where scientific publications should provide clear enough descriptions of the protocols to enable successful repetition and extension. Mesirov describes a *Reproducible Results System* that facilitates the enactment and publication of reproducible research. Such a system should provide the ability to track the provenance of data, analyses and results, and to package them for redistribution/publication. A key role of the publication is *argumentation*: convincing the reader that the conclusions presented do indeed follow from the evidence presented.

De Roure and Goble [5] observe that results are "reinforced by reproducibility", with traditional scholarly lifecycles focused on the need for *reproducibility*. They also argue for the primacy of method, ensuring that users can then reuse those methods in pursuing reproducibility. While traditional "paper" publications can present intellectual arguments, fostering reinforcement requires inclusion of data, methods and results in our publications, thus supporting reproducibility. A problem with traditional paper publications, as identified by Mons [6] is that of "Knowledge Burying". The results of an experiment are written up in a paper which is then published. Rather than explicitly including information in structured forms however, techniques such as text mining are then used to extract the knowledge from that paper, resulting in a loss of that knowledge.

In a paper from the Yale Law School Roundtable on Data and Code Sharing in Computational Science, Stodden et al. [7] also discuss the notion of Reproducible Research. Here they identify *verifiability* as a key factor, with the generation of verifiable knowledge being scientific discovery's central goal. They outline a number of guidelines or recommendations to facilitate the generation of reproducible results. These guidelines largely concern openness in the data publication process, for example the use of open licences and non-proprietary standards. Long term goals identified here include the development of version control systems for data; tools for effective download tracking of code and data in order to support citation and attribution; and the development of standardised terminologies and vocabularies for data description. Mechanisms for citation and attribution (including data citation, e.g. Data Cite[1]) are key in providing incentives for scientists to publish data.

The Scientific Knowledge Objects [8] of the LiquidPub project describe aggregation structures intended to describe scientific papers, books and journals. The approach explicitly considers the lifecycle of publications in terms of three "states": Gas, Liquid and Solid, which represent early, tentative and finalised work respectively.

Groth et al. [9] describe the notion of a "Nano-publication"—an explicit representation of a *statement* that is made in scientific literature. Such statements may be made in multiple locations, for example in different papers, and validation of that statement can only be done given the context. An example given is the statement that *malaria is transmitted by mosquitoes*, which will appear in many places in published literature, each occurrence potentially backed by differing evidence. Each nano-publication is associated with a set of annotations that refer to the statement and provide a minimum set of (community) agreed annotations that identify authorship, provenance, and so on. These annotations can then be used as the basis for review, citation and indeed further annotation. The Nano-publication model described in [9] considers a statement to be a *triple* – a tuple of three concepts, subject, predicate and object – which fits closely with the Resource Description Framework (RDF) data model [10], used widely for (meta)data publication (see the discussion on Linked Data below). The proposed implementation uses RDF and Named Graphs.[2] Aggregation of nano-publications will be facilitated by the use of common identifiers (following Linked Data principles as discussed in Section 7), and to support this, the Concept Web Alliance[3] are developing a ConceptWiki,[4] providing URIs for biomedical concepts. The nano-publication approach is rather "fine-grain", focusing on single statements along with their provenance.

The Executable Paper Grand Challenge[5] was a contest for proposals that will "improve the way scientific information is communicated and used". For executable papers, this will be through adaptations to existing publication models to include data and analyses and thus facilitate the validation, citation and tracking of that information. The three winning entries in 2011 highlight different aspects of the notion of executable papers. Collage [11] provides infrastructure which allows for the embedding of executable codes in papers. SHARE [12] focuses on the issue of reproducibility, using virtual machines to provide execution. Finally, Gavish and Donoh [13] focus on verifiability, through a system consisting of a Repository holding Verifiable Computational

[1] http://datacite.org/.
[2] See Section 7 for an explanation of Named Graphs.
[3] http://www.nbic.nl/about-nbic/affiliated-organisations/cwa/introduction/.
[4] http://conceptwiki.org/.
[5] http://www.executablepapers.com/.