Future Generation Computer Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

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 xxxx

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/.

Please cite this article in press as: S. Bechhofer, et al., Why linked data is not enough for scientists, Future Generation Computer Systems (2011), doi:10.1016/j.future.2011.08.004

Results (VCRs) that are identified using Verifiable Result Identifiers (VRIs). We note, however, that none of these proposals provide an explicit notion of "Research Object" as introduced here. In addition, provenance information is only considered in the third proposal, where Gavish and Donoh suggest that the ability to *re-execute* processes may be unnecessary. Rather, understanding of the process can be supported through providing access to the computation tree along with inputs, outputs, parameters and code descriptions.

2.2. Linked data

Benefits of explicit representation are clear. An association with a dataset (or service, or result collection, or instrument) should be more than just a citation or reference to that dataset (or service, or result collection). The association should rather be a *link* to that dataset (or service, or result collection, or instrument) which can be followed or dereferenced explicitly. Such linking provides access to the actual resource and thus enactment of the service, query or retrieval of data, and so on, fostering reproducibility.

The term Linked Data is used to refer to a set of best practices for publishing and connecting structured data on the Web [3]. Linked Data explicitly encourages the use of dereferenceable links as discussed above, and the Linked Data "principles" – use of HTTP URIs for naming, providing useful information when dereferencing URIs, and including links to other URIs – are intended to foster reuse, linkage and consumption of that data. Further discussion of Linked Data is given in Section 7.

2.3. Preservation and archiving

The Open Archival Information System (OAIS) reference model [14] describes "open archival information systems" which are concerned with preserving information for the benefit of a community. The OAIS Functional Model describes a core set of mechanisms which include Ingest, Storage and Access along with Planning, Data Management and Administration. There is also separation of *Submission Information Packages*, the mechanism by which content is submitted for ingest by a Producer; *Archival Information Package*, the version stored by the system; and *Dissemination Information Package*, the version delivered to a Consumer.

OAIS considers three external entities or actors that interact with the system. Producers, Management and Consumers, to characterise those who transfer information to the system for preservation; formulate and enforce high level policies (planning, defining scope, providing "guarantees") and are expected to use the information respectively. OAIS also considers a notion of a *Designated Community*, a subset of consumers that are expected to understand the archived information.

2.4. Scientific publication packages

One notable precursor to the notion of Research Object presented in this paper is the idea of *Scientific Publication Packages* (SPP), proposed in 2006 by Hunter to describe "the selective encapsulation of raw data, derived products, algorithms, software and textual publications" [15].

SPPs are motivated primarily by the need to create archives for the variety of artefacts, such as those listed above, that are produced during the course of a scientific investigation. In this "digital libraries" view of experimental science, SPPs ideally contain not only data, software, and documents, but their provenance as well. As we note here, the latter is a key enabler both for scientific reproducibility, and to let third parties verify scientific accuracy. Thus, SPPs are essentially containers that, unlike standard file packaging tools such as tar, or zip, adopt a specific terminology to provide a description of their content. Such terminology is an e-science specific extension of the ABC class hierarchy [16], previously proposed by the same authors as a generic taxonomy of terms for recording events in the lifecycle of digital objects in a library. Examples of specialisations include terms such as *Experiment* and *Simulation* (both types of *Event*), as well as *Model* and *Theory* (a type of *Work*). Although the taxonomy is simple and does not include terms to describe the *relationships* amongst the artefacts within a SPP, this proposal pre-dates the idea, common to our Research Objects, of combining containers with vocabularies for expressing a rich description of content.

To the best of our knowledge, the interesting preservation architecture designed around SPPs has remained at a prototype stage, making this more of an interesting point of reference than a baseline for a concrete implementation of an RO assembly and sharing toolkit.

2.5. Content vs. container

In terms of the conceptual models that can support the scientific process, there is much current interest in the representation of Scientific Discourse and the use of Semantic Web techniques to represent discourse structures (e.g. see [17]). Ontologies such as EXPO [18], OBI [19], MGED [20] and SWAN/SIOC [21] provide vocabularies that allow the description of experiments and the resources that are used within them. The HyPER community is focused on infrastructure to support Hypotheses, Evidence and Relationships. The Semantic Publishing and Referencing (SPAR) Ontologies[6] [22] also provide facilities for describing the component parts of documents and the scholarly publishing process.

In the main, however, this work tends to focus on the details of the relationships between the resources that are being described—what we might term the *content* rather than the *container*.

2.6. A motivating scenario

We use a scenario to motivate our approach and to illustrate aspects of the following discussion.

Alice runs an (in silico) analysis that involves the execution of a scientific workflow over some datasets. The output of the workflow includes results of the analysis along with provenance information detailing the services used, intermediate results, logs and final results. Outside the workflow she may add background information and interpretation of results. She collects together and publishes this information as an RO so that others can (1) validate that the results that Alice has obtained are fair; and (2) reuse the data, results and experimental method that Alice has described. Alice also includes within the RO links/mappings from data and resources used in her RO to public resources such as the ConceptWiki[7] or Linked Life Data,[8] providing additional context. Finally, Alice embeds the RO in a blog post so that others can access it.

Bob wants to reuse Alice's research results and thus needs sufficient information to be able to understand and interpret the RO that Alice has provided. Ideally, this should require little (if any) use of *backchannels*, direct or out-of-band communication with Alice. Bob can then deconstruct Alice's RO, construct a new experiment by, for example, replacing some data but keeping the same workflow, and then republishes on his blog, including in the new RO a link to Alice's original.

[6] http://opencitations.wordpress.com/2010/10/14/introducing-the-semantic-publishing-and-referencing-spar-ontologies/.

[7] http://conceptwiki.org/.

[8] http://linkedlifedata.com/.

Please cite this article in press as: S. Bechhofer, et al., Why linked data is not enough for scientists, Future Generation Computer Systems (2011), doi:10.1016/j.future.2011.08.004

The OAIS model considers three external entities or actors that may interact with the system, *producers*, *management* and *consumers*. In our scenario here, Alice is playing the role of producer, while Bob is a consumer. This desire to reduce the use of backchannels corresponds to the OAIS notion of [preserved] information being *independently understandable* in the sense that the information can be understood by users without the assistance of the information producer. Bob is thus a member of the *Designated Community* in OAIS terms.

In order to support this interaction, common structures for describing the resources and their relationships are needed. In addition, we require support for navigation/reference to external resources (such as ConceptWiki entries).

Importantly, ROs may contain references to data that is stored elsewhere. A number of data preservation initiatives are currently in place to ensure the long-term storage and reusability of scientific data on a large scale.[9] While this assumption pushes all data stewardship problems to dedicated data architectures, it also raises the new issue of resolving data references with no guarantee that the target has not been removed. In the RO model we take a best-effort approach that is similar to that of the Web architecture: there is indeed no guarantee that all links inside an RO can be resolved. On the other hand, unlike simple Web pages, ROs maintain a state, as described later in Section 6.3. Among other things, the state reflects the integrity of an RO with respect to the external resources it links to, at the time those resources are accessed.

2.7. Package, publish and preserve

We can identify at least three distinct processes or stages in the scenario described above.

Packaging. In conducting and describing her investigation, Alice brings together a number of different resources, for example a description of her hypothesis; datasets; workflows, scripts or analysis pipelines that she may have used to perform the investigation; intermediate and final results; and dissemination materials relating to the investigation, e.g. "the paper" (in a traditional sense) or presentation materials. These resources (some owned by Alice, some under the control of third parties) are brought together into a single package.

Publishing. Once materials are collected together, they can be exposed in a way that is then (re)usable by others. By publication here we refer to a process which involves the exposure or advertising of results. This could include aspects of "traditional" publication channels but is not limited to this. In the scenario described above, the embedding of an RO in Alice's blog is publication.

Preservation. Packaging and Publication make information available to others. Preservation aims to ensure that resources are made available in the future. Preservation may also require an element of *curation* and management of metadata or annotations relating to the preserved objects. In our scenario, once Bob has conducted his own investigation or experiment, making use of the resources and results packaged up in Alice's RO, he can repackage it along with any additional results of methods that he may have used into a new RO for archiving.

We explicitly consider here publication (exposure) as distinct from preservation. Although preservation is an important consideration in any approach supporting reusability, our focus here is chiefly on the packaging mechanisms (although see discussion of the Wf4Ever project in Section 8).

Fig. 1 shows some of the interactions involved in the scenario, along with the different stages identified above.

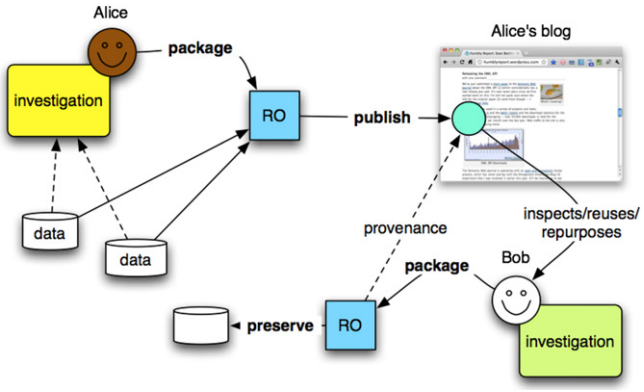


Fig. 1. User scenario.

2.8. Linked data is not enough!

Through the use of HTTP URIs and Web infrastructure, Linked Data provides a standardised publishing mechanism for structured data, with "follow your nose" navigation allowing exploration and gathering of external resources. For example, [23] uses a Linked Data approach to publish provenance information about workflow execution. The use of RDF (and thus associated representation machinery such as RDF Schema and OWL) offers the possibility of inference when retrieving and querying information.

What Linked Data does not explicitly provide, however, is a common model for describing the structure of our ROs and additional aspects that are needed in order to support the scholarly process—factors such as lifecycle, ownership, versioning and attribution. Linked Data thus says little about how that data might be organised, managed or consumed. Linked Data provides a platform for the sharing and publication of data, but simply publishing our data as Linked Data will not be sufficient to support and facilitate its reuse.

Jain et al. [24] also question the value of "vanilla" Linked Data in furthering and supporting the Semantic Web vision. Their concerns are somewhat different (although complementary) to ours here—with a focus on how one selects appropriate datasets from the "Linked Data Cloud", a concern about the lack of expressivity used in datasets (thus limiting the use to which reasoning can be usefully employed), and the lack of schema mappings between datasets. The nano-publications of Groth et al. [9] are also looking to add additional shared content on top of the Linked Data approach in terms of minimal annotations. Here we focus more on the need for a (common) aggregation model.

Note that this is not intended as a criticism of the Linked Data approach—simply an observation that additional structure and metadata is needed that sits on top of the Linked Data substrate and which then supports the interpretation and reuse of that data. Furthermore there is a need for the metadata to link the structure of the research resources with the function of the research process. A somewhat simplified picture is shown in Fig. 2 with the RO Layer providing a structured "view" on the underlying resources that can then be consumed by RO aware services.

What is missing, then, is a mechanism to describe the aggregation of resources, which through sufficient description of the contribution of these resources to the research and their relationships to each other, captures the additional value of the collection, and enables its reuse through the exchange of a single object. Scientific publication requires the representation of provenance, versioning, attribution, credit and the flow of intellectual rights.

Our notion of *Research Object* is intended to supply these aggregations and provide a container infrastructure, facilitating the

[9] One of these is the NSF-sponsored DataONE project (http://www.dataone.org), which caters to the Earth Sciences community and aims at preserving Observational data.

Please cite this article in press as: S. Bechhofer, et al., Why linked data is not enough for scientists, Future Generation Computer Systems (2011), doi:10.1016/j.future.2011.08.004

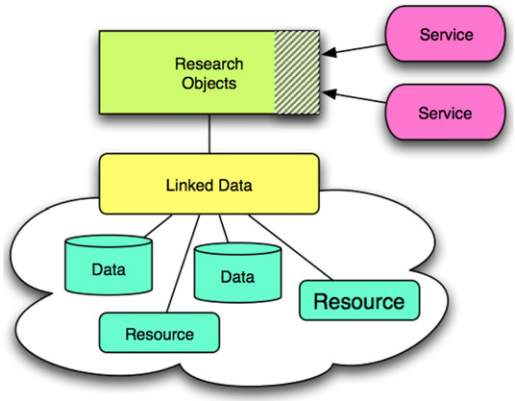


Fig. 2. Research object layer.

sharing and reuse of scientific data and results. Such a common model then facilitates the construction of services for the creation, manipulation and sharing of our research results.

3. Characterising reuse

In our scenario, we assert that Bob wants to reuse Alice's results and observe that the term "reuse" can be used to describe a range of activities. Reuse can come in many different forms, particularly when we consider reuse not just of data but also of method or approach. Thus an experiment or investigation may be *repeated*, enacting the same sequence of steps, or perhaps *repurposed*, taking an existing sequence of steps and substituting alternative data or methods in order to arrive at a new, derived, experiment. Objects can be reused as they can be decomposed and then recomposed in different ways. If they encapsulate processes, these processes can be re-enacted or previous executions of the process can be examined. As introduced above, *reproducibility* is key in supporting the validation of research.

Below, we introduce a number of principles intended to make explicit the distinctions between these kinds of general reuse, and identify the particular requirements that they make on any proposed e-Laboratory infrastructure.

We anticipate that due to the context of e-Laboratories, Research Objects will often encapsulate an enactable experiment or investigation. Thus some of our principles are driven by this assumption and refer in some way or other to being able to reuse or repeat the process.

Reusable. The key tenet of Research Objects is to support the sharing and reuse of data, methods and processes. Thus our Research Objects must be reusable as part of a new experiment or Research Object. By reuse here, we refer to a "black box" consideration of the Research Object where it is to be reused as a whole or single entity.

Repurposeable. Reuse of a Research Object may also involve the reuse of constituent parts of the Research Object, for example taking a study and substituting alternative services or data for those used in the study. By 'opening the lid' we find parts, and combinations of parts, available for reuse. The descriptions of the relationships between these parts and the way they are assembled are a clue as to how they can be reused. To facilitate such a disaggregation and recombination, Research Objects should expose their constituent pieces. Thus our Research Object framework also has need of an aggregation mechanism.

Repeatable. There should be sufficient information in a Research Object for the original researcher or others to be able to repeat the study, perhaps years later. Information concerning the services or processes used, their execution order and the provenance of the results will be needed. Repetition may involve access to data or execution of services, thus introducing a requirement for enactment services or infrastructure that can consume Research Objects. In the extreme, this may require, for example, virtual machines that recreate the original platform used to enact an analysis or simulation. In addition, the user will need sufficient privileges to access any data or services required.

Reproducible. To reproduce (or replicate) a result is for a third party to start with the same inputs and methods and see if a prior result can be confirmed. This can be seen as a special case of Repeatability where there is a complete set of information such that a final or intermediate result can be verified. In the process of repeating and especially in reproducing a study, we introduce the requirement for some form of comparability framework in order to ascertain whether we have indeed produced the same results. As discussed above, reproducibility is key in supporting the validation and non-repudiation of scientific claims.

Replayable. If studies are automated they might involve single investigations that happen in milliseconds or protracted processes that take months. Either way, the ability to replay the study, and to study parts of it, is essential for human understanding of what happened. Replay thus allows us to "go back and see what happened". Note that replay does not necessarily involve execution or enactment of processes or services. Thus replay places requirements on metadata recording the provenance of data and results, but does not necessarily require enactment services.

Referenceable. If ROs are to replace (or augment) traditional publication methods, then they (and their constituent components) must be referenceable or citeable. Thus mechanisms are needed for unambiguous reference to versions of ROs and which support discovery and retrieval.

Revealable. The issue of provenance, and being able to audit experiments and investigations is key to the scientific method. Third parties must be able to audit the steps performed in the research in order to be convinced of the validity of results. Audit is required not just for regulatory purposes, but allows for results to be interpreted and reused. Thus an RO should provide sufficient information to support audit of the aggregation as a whole, its constituent parts, and any process that it may encapsulate.

Respectful. A key aspect of e-Laboratories is user-visibility, credit and attribution. The paper citation count is an important metric in measuring the visibility and impact of published work. If we move to RO based publishing, we will require a re-engineering of reward structures for scientists—citation counts are no longer enough if derived works are being built through the reuse or repurposing of data and methods. Explicit representations of the provenance, lineage and flow of intellectual property associated with an investigation are needed.

4. RO principles, behaviours and features

The main purpose of Research Objects is to provide a class of artefacts that can encapsulate digital knowledge and provide a mechanism for sharing and discovering assets of reusable research and scientific knowledge.

The variety of reusabilities can be seen as a collection of behaviours that we expect our shareable objects to exhibit—these then place requirements on the ways in which our models are defined, and this in turn informs the features of the Research Object Model and the services that will produce, consume and manipulate ROs.

The principles stated above describe properties or constraints on the way in which we see ROs being used or behaving. Below, we outline a number of features that can facilitate the delivery of this functionality.

Aggregation. ROs are aggregations of content. Aggregation should not necessarily duplicate resources, but allow for references

Please cite this article in press as: S. Bechhofer, et al., Why linked data is not enough for scientists, Future Generation Computer Systems (2011), doi:10.1016/j.future.2011.08.004

to resources that can be resolved dynamically. There may also, however, be situations where, for reasons of efficiency or in order to support persistence, ROs should also be able to aggregate literal data as well as references to data.

Identity. Fundamental to Information Retrieval Systems is the ability to refer uniquely to an object instance or record by an identifier that is guaranteed to be unique throughout the system in which it is used. Such mechanisms must allow reference to the Object as a whole as well as to the constituent pieces of the aggregation. Identity brings with it the requirement for an account of equivalence or equality. When should objects be considered equivalent? Alternatively, when can one object be substituted for another? This will be context dependent; for example, in a given context, two objects may not be considered equivalent, but may be substitutable (e.g. either could be used with the same results).

Metadata. Our e-Laboratory and RO framework is grounded in the provision of machine readable and processable metadata. ROs will be annotated as individual objects, while metadata will also be used to describe the internal structures and relationships contained within an RO. Metadata can describe a variety of aspects of the RO, from general "Dublin Core" style annotations through licensing, attribution, credit or copyright information to rich descriptions of provenance or the derivation of results. The presence of metadata is what lifts the RO from a simple aggregation (e.g. a zip file) to a reusable object.

Lifecycle. The processes and investigations that we wish to capture in the e-Laboratory have a temporal dimension. Events happen in a particular sequence, and there are lifecycles that describe the various states through which a study passes. ROs have state, and this state may impact on available operations. For example, a study may go through a number of stages including ethical approval, data collection, data cleaning, data analysis, peer review and publication. At each stage in the process, it may be possible to perform different actions on the object. Thus a principled description of RO lifecycle is needed in our framework (see Section 6).

Versioning. In tandem with Lifecycle comes Versioning. ROs are dynamic in that their contents can change and be changed. Contents may be added to aggregations, additional metadata can be asserted about contents or relationships between content items and the resources that are aggregated can change. ROs can also be historical, in that they capture a record of a process that has been enacted. Thus there is a need for versioning, allowing the recording of changes to objects, potentially along with facilities for retrieving objects or aggregated elements at particular points in their lifecycle (see Section 6).

Management. The management of ROs will require operations for Creation, Retrieval, Update, Deletion (CRUD) of those objects. Storage is also a consideration.

Security. ROs are seen as a mechanism to facilitate sharing of data, methods and expert guidance and interpretation. With sharing come issues of access, authentication, ownership, and trust that we can loosely classify as being relevant to Security.

Graceful Degradation of Understanding. Finally, we outline a principle that we believe is important in delivering interoperability between services and which will aid in reuse of ROs, particularly serendipitous or unpredicted reuse—"Graceful Degradation of Understanding". RO services should be able to consume ROs without necessarily understanding or processing all of their content. ROs contain information which may be domain specific (for example, properties describing relationships between data sources and transformations in an investigation). Services should be able to operate with such ROs without necessarily having to understand all of the internal structure and relationships. This places a requirement of principled extensibility on the RO model.

This notion of Graceful Degradation of Understanding can also be observed in the layering approach used, for example, in Semantic Web representation languages. Triple store infrastructure can be used to store data represented using RDF graphs. Such graphs may include the use of vocabularies or representations—for instance, descriptions could be applied to resources making use of OWL [25] ontologies. The underlying triple store does not necessarily need to "understand" the semantics of OWL in order to provide useful functionality. For example a number of triple stores support hierarchical classification using simple RDF(S) [26] reasoning. Of course, if applications do understand upper layers, they can provide additional functionality or services.

5. Representation and implementation

In practice, during the lifecycle of an investigation (which spans activities including planning, execution of experiments or gathering of observational data, analysis of data and dissemination/publication) scientists will work with multiple content types with data distributed in multiple locations. Thus scientists use a plethora of disparate and heterogeneous digital resources. Although potentially useful individually, when considered collectively these resources enrich and support each other and constitute a scientific investigation [27].

These resources may vary widely depending on domain, discipline and the particular investigations being performed. We can, however, identify how individual resources constitute familiar parts of an investigation, and these are among the pieces that will make up our ROs.

Questions around a research problem, with or without a formal hypothesis. Descriptions or abstracts.

Organisational context. Ethical and governance approvals, investigators, etc. Acknowledgements.

Study design encoded in structured documents,

Methods scientific workflows or scripts, services, software packages.

Data from observations or measurements organised as input datasets.

Results from analyses or in silico experiments. Observations, derived datasets, along with information about their derivation or capture—provenance, algorithms, analyses, instrument calibrations.

Answers. Publications, papers, reports, slide-decks, DOIs, PUBMED ids etc.

A number of different projects have already been developing what one might describe as RO frameworks. These projects are "e-Laboratories" — environments providing a distributed and collaborative space for e-Science, enabling the planning and execution of in silico and hybrid experiments; i.e. processes that combine data with computational activities to yield experimental results.

Here we discuss this work and how it relates to our overall vision of ROs.

5.1. myExperiment

The myExperiment Virtual Research Environment[10] has successfully adopted a Web 2.0 approach in delivering a social web site where scientists can discover, publish and curate scientific workflows and other artefacts. While it shares many characteristics with other Web 2.0 sites, myExperiment's distinctive features to meet the needs of its research user base include support for

[10] http://www.myexperiment.org.

Please cite this article in press as: S. Bechhofer, et al., Why linked data is not enough for scientists, Future Generation Computer Systems (2011), doi:10.1016/j.future.2011.08.004

credit, attributions and licensing, and fine control over privacy. myExperiment now has around 3000 registered users, with thousands more downloading public content, and the largest public collection of workflows. Over time, myExperiment has embraced several workflow systems including the widely-used open source Taverna Workflow Workbench. Created in close collaboration with its research users, myExperiment gives important insights into emerging research practice.

In terms of our reuse characterisations, simply sharing workflows provides support for *repurposing*, in that workflows can be edited, and re-run. myExperiment recognised [28] that workflows can be enriched through a bundling of the workflow with additional information (e.g. input data, results, logs, publications) which then facilitates *reproducible* research. In myExperiment this is supported through the notion of "Packs", collections of items that can be shared as a single entity.

The pack allows for basic aggregation of resources, and the pack is now a single entity that can be annotated or shared. In order to support more complex forms of reuse (for example, to rerun an investigation with new data, or validate that the results being presented are indeed the results expected), what is needed in addition to the basic aggregation structure, is metadata that describes the relationships between the resources within the aggregation. This is precisely the structure that ROs are intended to supply, the basic pack aggregation being enhanced through the addition of metadata capturing the relationships between the resources—for example the fact that a particular data item was produced by the execution of a particular workflow. The pack (or RO) then provides a context within which statements can be made concerning the relationships between the resources. Note that this is then one viewpoint—other ROs could state different points of view regarding the relationships between the (same) resources in the RO. We return to a discussion of representation in myExperiment in Section 7.

5.2. SysMO SEEK

Systems Biology of Microorganisms (SysMO)[11] is a European trans-national research initiative, consisting of 91 institutes organised into thirteen projects whose goal is to create computerised mathematical models of the dynamic molecular processes occurring in microorganisms. SysMO-DB[12] is a web-based platform for the dissemination of the results between SysMO projects and to the wider scientific community. SysMO-DB facilitates the web-based exchange of data, models and processes, facilitating sharing of best practice between research groups.

SysMO SEEK[13] is an "assets catalogue" describing data, models, Standard Operating Procedures (SOPs), and experiment descriptions. Yellow Pages provide directories of the people who are involved in the project.

SysMO SEEK provides a retrospective attempt to share data and results of investigation along with the methods that were used in their production. The implementation is built upon, and specialises, generic components taken from the myExperiment project.

A number of challenges characterise SysMO-SEEK. Users want to keep their current, bespoke data formats, with a significant support for spreadsheets. Consequently, individual projects are responsible for keeping their own data in separate repositories requiring a framework which allows for references to data that can be resolved upon request.

Projects are also cautious about data access, sharing and attribution, resulting in a sophisticated model of sharing and access control where data and models can be shared with named individuals, groups, projects, or the whole community at the discretion of the scientists. This supports not only the publication of results, but also collaborative working and sharing.

The information in SysMO SEEK is structured using a model called JERM (Just Enough Results Model) which allows the exchange, interpretation and comparison of different types of data and results files across SysMO. JERM is based on the ISA (Investigation/Study/Assay) [29] format. Within SysMO, experiments are described as Assays, which are individual experiments as part of a larger Study. These Studies themselves are part of a much larger Investigation. The aim is that the JERM will move towards linking Models (Biological models, such as SBML) together with the experimental data that was used to both construct and test the model, within the context of one or more Assays. The JERM model extends ISA and provides specific relationships appropriate to the domain. The ISA format is, however, somewhat "top down", allowing for the packaging of data relating to specific investigations or studies, but less appropriate for re-assembling or reusing data for differing sources.

ROs would then encapsulate the Model together with information about its simulation environment, parameters and data thereby providing a third party with everything they need to reproduce and validate the model, along with the hypothesis and provenance behind its creation. An addition, this description of the *Experimental Narrative* is a feature that we are likely to see needed in other scenarios.

In the Systems Biology community, the requirement for ROs has already been recognised. Emerging standards and markup languages, such as SBRML (Systems Biology Results Markup Language) [30] extend the SBML model format to allow scientists to encapsulate experimental data links with their models. This allows both the representation of computational simulation results and experimental results in the context of a particular model.

Returning to our characterisation of reuse, many of the processes currently described within SysMO are actually wet-lab experiments. As a result, *traceability* and *referenceability* are the key kinds of reuse that are needed within SysMO, allowing for validation of the results. With greater use of workflows in the future, *repeatability* and *replayability* will begin to play a part.

5.3. MethodBox and NHS e-Lab

MethodBox[14] is an environment for finding variables from data archives for cross disciplinary research effectively "turning data archives into data playgrounds". It is part of the Obesity e-Lab project [31] addressing the public health need for greater understanding of obesity across biomedical and social perspectives, where researchers from different disciplines may use common data archives but do not usually cross-fertilize their work. The generic MethodBox environment is built on the concept of a social network of researchers "shopping for variables". A variable is a vector of data measured or observed in some way about a factor such as age, sex, body mass index, etc. The elements of the vector usually relate to an individual taking part in a study. Archives such as the UK Data Archive[15] contain millions of variables grouped into sets such as annual surveys, for example the Health Surveys for England. The supporting documentation for each survey typically contains important metadata about variables. Researchers

11 http://www.sysmo.net/.
12 http://www.sysmo-db.org/.
13 http://www.sysmo-db.org/seek/.
14 http://www.methodbox.org/.
15 http://www.data-archive.ac.uk/.

Please cite this article in press as: S. Bechhofer, et al., Why linked data is not enough for scientists, Future Generation Computer Systems (2011), doi:10.1016/j.future.2011.08.004

may take days to wade through supporting documentation and large datasets to extract the variables and metadata they need. Methodbox reduces the time required from days to minutes. It does this by mapping variables to metadata from: (1) relevant parts of supporting documentation; (2) sets of variables extracted by users; (3) user-contributed scripts and guidance for deriving, transforming or using variables. A derived variable might be a categorisation of socio-economic status based on household income and other factors, or a categorisation of obesity based on body mass index. The social scientist may have more expertise in measuring socio-economic status and the biomedical researchers expertise focuses on obesity. Thus a cross-talk between disciplines may emerge at a very early stage of research by sharing methods applicable to variables of common interest. Users are able to share their expertise over particular survey variables such as the way questionnaire responses about smoking can be made "research ready" and then analysed appropriately. Scripts for extracting sets of variables, transforming multiple variables into one and building research models are the currency of sharing. The sharing of scripts leads to *repurposing* of study methods.

The sets of variables in MethodBox "shopping baskets" may be seen as incomplete ROs intended to seed ROs for analysis in external e-Laboratories. In addition, ROs may be initiated elsewhere before being populated with data preparation methods and data extracts in MethodBox. So Methodbox is taking the "Research Object on the inside" approach, anticipating future value of reuse and audit of the semantic aggregation of research entities.

NHS e-Lab[16] is an e-Laboratory for socially networked "sense-making" over health related datasets. It operates within the UK National Health Service firewall and introduces the notion of a federation of e-Laboratories sharing ROs across organisational boundaries after checks that the RO does not contain material that might identify a patient. In addition to security, there is a strong requirement to increase the consistency and efficiency with which NHS analysts perform analyses. The *repurposing* of ROs encourages sharing of templates instead of duplication of similar analytical processes; the *revealability* enhances information governance; the *repeatability* builds organisational memory; and the *respectfulness* helps to build a reward environment, which can be linked to continuing professional development credits.

In order to "borrow strength" from academia, NHS e-Lab is designed to import ROs from MethodBox where national survey data is needed by those planning local NHS services. Attribution, sharing and audit logs will become particularly important for cross organisation as well as cross discipline sharing.

6. RO stereotypes and versioning

In this section we characterise ROs in terms of a small number of *stereotypes*, i.e., common patterns of resource aggregation that emerge from an examination of our projects involved in e-Laboratory related activities. Stereotypes characterise ROs according to the two orthogonal dimensions of *state* and *purpose*.

More specifically, we introduce *lifecycle* stereotypes, describing states that ROs can transition into and out of as part of their evolution in time (Section 4), and *functional* stereotypes which describe the role of the RO in the context of data sharing. We then describe RO evolution in terms of updates and versioning operations that affect state.

6.1. Lifecycle stereotypes

Live objects (LO) represent a work in progress. They are thus mutable as the content or state of their resources may change, leading to the need for version management. LOs are potentially under the control of multiple owners and may fall under mixed stewardship, raising issues of security and access control.

Publication objects (PO) are intended as a record of past activity, ready to be disseminated as a whole. This is in line with our key motivation for ROs, namely to support "rich publication" by moving from traditional paper based (linear) dissemination mechanisms, to aggregations of related and interlinked pieces of information. POs are immutable, and their multiple successive versions are considered as distinct objects. They must be citeable, and credit and attribution are central aspects of the publication process as they are key to providing rewards, and thus incentives, for scientific publication. POs may also make use of ontologies for the representation of the rhetorical or argumentation structure in the publication (see Section 2.5).

Archived objects (AO) encapsulate aggregations that represent the endpoint of an RO's life, either because it is now deprecated, or has reached a version that the author prescribes to be final. AOs are therefore *immutable*, with no further changes or versions allowed. For example, an AO may represent a historical record for resources used in an experiment which has concluded, or has been abandoned.

With this simple state classification, we can describe the lifetime of an RO in terms of its evolution from LO, to either PO or AO (the "terminal states"), while at the same time multiple versions of an LO may be created, each evolving independently into POs or AOs.

6.2. Functional stereotypes

Work objects encompass ROs and extend the applications beyond research, for example to business intelligence and audit—where repeatability, replayability and repurposing are key aspects [1].

Exposing objects are wrappers that provide a standardised metadata container for existing data. For example, spreadsheets may be gathered together and aggregated along with the methods used to produce them. This aggregation can be seen as an RO, but it can also be a smaller component, exposing the spreadsheet collection to the RO thereby setting it in a reproducible research context. The Exposing Object provides a Wrapper [32] that allows the spreadsheet to be seen as an RO, facilitating its exposure and integration into the Web of Linked Data.

View/context objects can provide a view over some already exposed data. It is here that ROs can interact with data that is exposed or published using Linked Data principles [3], providing a "Named Graph" for those resources.

Method objects contain methods and descriptions of methods—enabling methodological research to be exposed in an RO and consumed by other ROs in applied, as distinct from methodological, research. This may help to propagate methodological integrity and avoid translation errors for methods.

The OAIS model [14] also identifies variants of aggregation such as *dissemination* and *archival* information packages, corresponding loosely to our notion of publication or archived objects.

6.3. Evolution and versioning

At any given point in time, an RO is characterised by (i) its lifecycle stereotype, defined earlier; (ii) its version, and (iii) its value, defined as the union of the values of all its components. Note that when internal components are known by reference, i.e., via

[16] http://www.nweh.org.uk/.

their URIs or other Open Data links, the value of the referenced content is represented by the value of the reference. This means that the value of an RO does not change if an update to any of its components is not reflected by providing a new reference to it.

RO evolution may occur in three possible ways:

by state transition: the legal transitions involving the lifecycle stereotypes are shown in Fig. 3(a);
by in-place update: the *update* operation produces a new value for the same RO and retains its identity;
by versioning: the *versioning* operation produces a new RO with a new identity and a new value.

Fig. 3(b) shows a possible evolution of a Live Object X. Version 1 of X, denoted X_1(LO), is updated multiple times prior to being published at time t_3, as X_1(PO). A new version X_2(LO) of X_1(LO) is then created, which is itself updated multiple times, prior to being archived as X_2(AO). Independently, X_1(PO) is also archived at some other time t_5. Note that, according to the state diagram, neither X_1 nor X_2 can evolve further after reaching their AO state.

A key characteristic relates to the (im)mutability of both the resources described and the relationships between them. Neither Linked Data nor OAI-ORE tackle the issue of versioning explicitly. In the case of an Archived Object, when a scientist returns to it, it should refer to the same versions of the data that were originally used. Live Objects, however may have contents that are updated.

Mechanisms such as the Probity service[17] that allow a client to detect when changes have occurred have a role to play here as does the Memento framework of Van de Sompel [33]. Probity allows for recordings of "checksums", providing some minimum guarantees as to whether resources have changed (but not necessarily providing solutions as to what to do when changes have occurred). Memento provides a versioning mechanism for resources that allows "follow your nose" style navigation as widely used in the Linked Data approach. Considering OAIS again, we can see that our Archive Objects relate to Archival Information Packages which should contain information relating to the preservation (provenance, fixity, etc.).

7. Implementing ROs: linked data and OAI-ORE

Although our argument here is that Linked Data alone is not enough to support scientific publication (and the reuse, reproduction or validation of results), Linked Data does offer a substrate for data publication that can then be used by additional layers providing the necessary aggregations, provenance, attribution, etc.

7.1. Linked data

The term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web [3], intended to foster reuse, linkage and consumption of that data. The principles can be summarised as follows:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up or dereferences a URI, provide useful information, using standard representations (for example RDF).
4. Within that useful information, include links to other URIs. so that clients can discover more things.

In the five years or so since the first discussions of the Linked Data approach, the amount of linked data published has been increasing rapidly. The Open Data movement has seen successful pressure on governments to expose and open up data sets—in many cases this is being done using a Linked Data approach. An example of this is data published by the UK Ordnance Survey,[18] which provides a SPARQL [34] endpoint (allowing query against an RDF triple store) to data describing administrative geography in the UK. Similar government initiatives are also in place in other countries including the US.

Within the scientific community, datasets are also being exposed using a Linked Data approach. Bio2RDF [35] provides "rdfized" access to information from data sources such as Kegg, PDB, MGI and HGNC. The Linking Open Drug Data (LODD)[19] activity of W3C's Health Care and Life Sciences Interest Group is publishing and interlinking information relating to drugs using Linked Data. Linked Life Data[20] provides a platform for integrating a number of data sources including UniProt, UMLS, Drug information, PubMed. Other sources exposed as Linked data include species data,[21] Clinical Trials,[22] MeSH[23] and the ConceptWiki[24] as discussed above.

Our intention is that the basic concept of ROs should be independent of the mechanism used to represent and deliver those objects. However the Linked Data approach has a good fit with the notion of ROs.

Within a (semantic) Web context, the term information resource is used to distinguish those resources (things that might be identified by URIs) for which it is the case that their essential characteristics can be conveyed in a message. Non-information resources are those things that might be identified by URIs, but for which this is not the case. Thus web pages, PDF documents, JPG images are examples of information resources, while people are non-information resources. A number of patterns have been identified [36] using techniques such as content negotiation and HTTP redirection, that support the description of non-information resources.

Thus the separation of the identity of an RO from serialisations of the description of its content reflects the handling of non-information resources—we consider a particular RO to be a non-information resource which may have alternative concrete representations. See below for further discussion of the use of non-information resources within myExperiment.

7.2. Aggregation

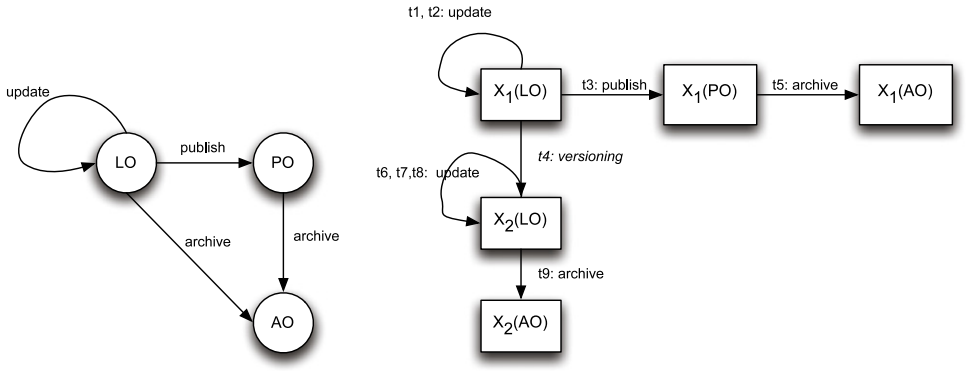
The idea of aggregation in a web context has already been addressed by the Open Archives Initiation Object Reuse and Exchange Specification (OAI-ORE, or ORE [37]). ORE defines a data model and a number of concrete serialisations (RDF, Atom and RDFa) that allow for the description of aggregations of Web resources. The key concepts in ORE are the notions of Aggregation, which represents an aggregation of a number of resources; and ResourceMap, which provides a concrete representation of the elements in the aggregation (AggregatedResources) and relationships between them.

The ORE model is agnostic as to the semantics of such aggregations—examples are given which include aggregations of

[17] http://www.probity.org/.

[18] http://data.ordnancesurvey.co.uk/.
[19] http://www.w3.org/wiki/HCLSIG/LODD.
[20] http://linkedlifedata.com/.
[21] http://lod.geospecies.org/ and http://www.taxonconcept.org/.
[22] http://linkedct.org/.
[23] http://www.nlm.nih.gov/mesh/.
[24] http://conceptwiki.org/.

Please cite this article in press as: S. Bechhofer, et al., Why linked data is not enough for scientists, Future Generation Computer Systems (2011), doi:10.1016/j.future.2011.08.004



(a) RO state transitions. (b) Possible evolution of RO instance X.

Fig. 3. Research object state transition and an example of object evolution.

favourite images from Web sites, the aggregation of a number of different resources to make up a publication in a repository, or multi-page HTML documents linked with "previous" and "next" links.

ORE provides a description of Resource Map Implementations using RDF [38], which integrates well with current approaches towards the publication of Linked Data [21].

7.2.1. Aggregations in myExperiment

Work in myExperiment makes use of the OAI-ORE vocabulary and model in order to deliver ROs in a Linked Data friendly way [39]. Although specific to myExperiment, the following discussion is pertinent to the other e-Laboratories.

In myExperiment, *packs* are created using a shopping basket (or wish list) metaphor. Typical packs contain workflows, example input and output data, results, logs, PDFs of papers and slides. To explore the extension of packs to richer ROs a service has been deployed which makes myExperiment content available in a variety of formats. Following "Cool URI" guidelines,[25] entities in myExperiment are considered as Non-Information Resources and they are given URIs. Content negotiation is then used to provide appropriate representations for requests, separating the resources from their explicit representations. RDF metadata is published according to the myExperiment data model which uses a modularised ontology drawing on Dublin Core, FOAF, OAI-ORE, SWAN-SIOC, Science Collaboration Framework, and the Open Provenance Model (OPM[26]) [40]. In addition to this "Linked Data" publishing, myExperiment content is also available through a SPARQL endpoint[27] and this has become the subject of significant interest within the community. It is effectively a generic API whereby the user can specify exactly what information they want to send and what they expect back—rather than providing query/access mechanism via specific API functions. In some ways it has the versatility of querying the myExperiment database directly, but with the significant benefit of a common data model which is independent of the codebase, and through use of OWL and RDF it is immediately interoperable with available tooling. Exposing data in this way is an example of the "cooperate don't control" principle of Web 2.0.

This brings myExperiment into the fold of the other SPARQL endpoints in e-Science, especially in the life sciences area [27] and we are beginning to see workflows that use the data provided by such endpoints. In minutes a user can assemble a pipeline which integrates data and calls upon a variety of services from search and computation to visualisation. While the linked data movement has persuaded public data providers to deliver RDF, we are beginning to see assembly of scripts and workflows that consume it — and the sharing of these on myExperiment. We believe this is an important glimpse of future research practice: the ability to assemble with ease experiments that are producing and consuming this form of rich scientific content.

7.2.2. Extended aggregation vocabularies

Publishing the myExperiment data using Linked Data principles facilitates the consumption of that data in applications, but needs further shared infrastructure to support the description of the RO structure. An RO is essentially an aggregation of resources, and we are using ORE as the basis for describing our RO. As we have mentioned, however, ORE only provides a general vocabulary for describing the relationships between resources in terms of an aggregation, but says nothing about the particular semantics of the relationships between resources. Thus there is no way, for example, of distinguishing between an aggregation of resources in a publication, and the constituent pages in a multi-page HTML document. To enable the description of specific relationships between the aggregated resources, the set of ORE relationships must be extended.

We consider two such extensions here. Firstly, the *Research Objects Upper Model* (ROUM) provides basic vocabulary that is used to describe general properties of RO that can be shared across generic e-Laboratory services. For example, the basic lifecycle states of ROs (as described in Section 4) are described in this upper model. Secondly, *Research Object Domain Schemas* (RODS) provide application or domain specific vocabulary for use in RO descriptions. For example, an RO may contain a reference to a service and a data item, along with an assertion that the data was produced through an invocation of the service. Applications which are aware of the intended semantics of the vocabulary used for these assertions can exhibit appropriate behaviour. It is important to stress here that applications that are not aware of these vocabularies will still be able to operate on the overall aggregation structure. This layered approach therefore helps meet our principle for Graceful Degradation of Understanding across e-Laboratory services (see Section 4). OAI-ORE has also been used in other efforts aimed at providing aggregations over scientific data such as SCOPE [41].

The interaction with a Linked Data view of the world is two-fold here. Firstly, one could view the RO as "Named Graphs for Linked Data", through the definition of an explicit container. The concept of Named Graphs in Semantic Web architecture and languages allows for the identification of selected subgraphs within a single triple or RDF graph. This helps to overcome some of the difficulties

25 http://www.w3.org/TR/cooluris/.
26 http://openprovenance.org/.
27 http://rdf.myexperiment.org.

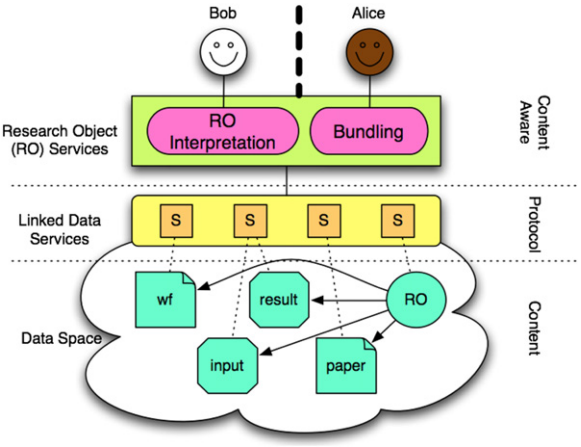


Fig. 4. Detailed layers.

in using a simple triple model, for example by being able to assert provenance or trust information with a particular collection of statements in a graph without resorting to reification. Named graphs extend the RDF data model and are supported in query languages such as SPARQL [34].

This also facilitates the exposure or publication of digital content as linked data. Secondly, the RO may also be a *consumer* of linked data, with linked data resources being aggregated within it.

Fig. 4 shows an enriched view of the layers presented earlier, following a common pattern of exposing *content* through a *protocol* layer to a collection of *content aware* services. The Linked Data Services provide a common mechanism exposing resources in the Data Space. The common protocols adopted here (use of Web architecture, HTTP, URIs, etc.) facilitate access to those resources, but are agnostic as to the content of those resources. Research Object services support the bundling together of resources into ROs (ingest) and the subsequent interpretation of those ROs (access).

8. Discussion

This paper sets out what can be seen as a manifesto for the Research Object concept and approach. We have discussed ways in which information can be repurposed, reused or validated in order to support the scientific research process. These ideas are currently being pursued in a number of research projects.

A challenge common to all emerging collaborative environments that promote open science and the rapid exchange of experimental and pre-publication data and methods is one of trust. As an identifiable container, Research Objects allow us to attribute a measure of trust to the object itself [42], with potential to apply and extend methods for modelling and computing social trust [43], trust in content [44] and trust based on provenance information [45].

The provision of reproducible results requires more than traditional paper publication—or even electronic publication but following the "paper metaphor". Linked Data provides some of the infrastructure that will support the exposure and publication of data and results, but will not alone enable reusable, shared research and the reproducibility required of scientific publication. Additional mechanisms are needed that will allow us to share, exchange and reuse digital knowledge as (de)composable entities. Our solution to this is ROs, semantically rich aggregations of resources that bring together the data, methods and people involved in (scientific) investigations.

The RO concept provides a layer of aggregation structure that is consistent with the Linked Data view of the world. ROs are both: (1) resources accessible via linked data principles; and (2) will aggregate linked data resources.

As discussed in Section 2.4, previous work has defined the notion of the Scientific Publishing Packages (SPP). Where SPPs diverge from ROs is more in the *intent* than in the *structure*: while SPPs are essentially designed for archival purposes, the lifecycle of a Research Object is instead centred around the notion of partial sharing, reusing, and possibly repurposing, making the issue of self-consistency of an RO central to our model.

A number of existing projects are already beginning to apply the RO approach to organising and publishing their data. In particular, myExperiment and NHS e-Lab have notions of prototypical ROs, and the capability to export them using Linked Data principles. By reflecting on how such aggregations play a part in the scientific process, we have proposed a set of principles and features.

Our next steps are to further refine these principles and features and provide implementations that support the lifecycle stages as identified here. The Wf4Ever ("Workflow for Ever") project,[28] aims to support the preservation and efficient retrieval and reuse of scientific workflows. The RO approach is central to Wf4Ever, with workflows being the prime content of the ROs generated. The ROs will be used as containers to package together workflows with data, results and provenance trails, with a key consideration being to support *preservation* of results. Approaches for validating integrity and mitigating against workflow *decay* are particular areas of interest for the project—this introduces requirements for aspects such as Aggregation, Versioning and Lifecycle as discussed in Section 4 and will allow us to further investigate issues of preservation which are not explicitly considered here. Two contrasting domains are being explored, Astronomy and Genomics—in the initial use cases, validation and verification of results will be the focus. Current explorations include investigation into more detailed stereotypes and the production of a "Research Object Zoo" identifying concrete examples of the broad classifications introduced in Section 6. In addition, particular use case scenarios within those domains are being used to identify different user roles and their interactions with Research Objects at different states in the lifecycle as discussed in 6; and identify more specific requirements for content of Research Objects and the vocabularies needed to describe relationships between that content. Use cases cover varying scenarios include the analysis of existing gene expression data from wet lab experiments (genomics) and calculation of luminosities for galaxies.

In closing, we believe that the RO approach will enable us to conduct scientific research in ways that are: efficient, typically costing less to borrow a model than create it; effective, supporting larger scale and deeper research by reusing parts of models; and ethical, maximising benefits for the wider community, not just individual scientists, with publicly funded research.

Acknowledgments

This paper reflects the experiences within several e-Laboratory projects represented by the authors. We wish to thank the project funders and the many other members of the project teams who have contributed indirectly to this work.

References

[1] J. Ainsworth, I. Buchan, e-Labs and work objects: towards digital health economies, in: Comms. Infrastructure. Systems and Applications in Europe, vol. 16, 2009, pp. 206–216.
[2] G. Olson, A. Zimmerman, N. Bos, Scientific Collaboration on the Internet, MIT Press, 2008.

[28] http://www.wf4ever-project.org/ funded under FP7's Digital Libraries and Digital Preservation Call. (ICT-2009.4.1).

[3] C. Bizer, T. Heath, T. Berners-Lee, Linked data—the story so far, International Journal on Semantic Web and Information Systems 5 (3) (2009) 1–22.
[4] J.P. Mesirov, Accessible reproducible research, Science 327 (5964) (2010) 415–416.
[5] D. De Roure, C. Goble, Anchors in shifting sand: the primacy of method in the web of data, in: Web Science Conference 2010, Raleigh, NC, 2010.
[6] B. Mons, Which gene did you mean? BMC Bioinformatics 6 (2005) 142.
[7] Yale Roundtable Participants, Reproducible research: addressing the need for data and code sharing in computational science, Journal of Computing Science and Engineering 12 (5) (2010) 8–13. doi:10.1109/MCSE.2010.113.
[8] F. Giunchiglia, R. ChenuAbente, Scientific knowledge objects V.1, Technical Report DISI-09-006, University of Trento, January 2009.
[9] P. Groth, A. Gibson, J. Velterop, The anatomy of a nano-publication, Information Services and Use 30 (1) (2010) 51–56. URL: http://iospress.metapress.com/index/FTKH21Q50T521WM2.pdf.
[10] G. Klyne, J.J. Carroll, Resource description framework (RDF): concepts and abstract syntax, W3C Recommendation, World Wide Web Consortium, 2004. URL: http://www.w3.org/TR/owl-guide/.
[11] P. Nowakowski, E. Ciepiela, D. Harezlak, J. Kocot, M. Kasztelnik, T. Bartynski, J. Meizner, G. Dyk, M. Malawski, The collage authoring environment, in: Proceedings of the International Conference on Computational Science, ICCS 2011, Procedia Computer Science 4 (2011) 608–617. doi:10.1016/j.procs.2011.04.064.
[12] P.V. Gorp, S. Mazanek, Share: a web portal for creating and sharing executable research papers, in: Proceedings of the International Conference on Computational Science, ICCS 2011, Procedia Computer Science 4 (2011) 589–597. doi:10.1016/j.procs.2011.04.062.
[13] M. Gavish, D. Donoho, A universal identifier for computational results, in: Proceedings of the International Conference on Computational Science, ICCS 2011, Procedia Computer Science 4 (2011) 637–647. doi:10.1016/j.procs.2011.04.067.
[14] Consultative Committee for Space Data Systems, Reference model for an open archival information system (OAIS), Blue Book CCDS 650.0-B-1, Open Archives Initiative, 2002.
[15] J. Hunter, Scientific publication packages—a selective approach to the communication and archival of scientific output, International Journal of Digital Curation 1 (1) (2006). URL: http://www.ijdc.net/index.php/ijdc/article/view/8.
[16] C. Lagoze, J. Hunter, The ABC ontology and model, Journal of Digital Information 2 (2) (2002). URL: http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Lagoze/.
[17] T. Clark, J.S. Luciano, M.S. Marshall, E. Prud'hommeaux, S. Stephens (Eds.), Semantic Web Applications in Scientific Discourse 2009, vol. 523, in: CUER Workshop Proceedings, 2009.
[18] L.N. Soldatova, R.D. King, An ontology of scientific experiments, Journal of the Royal Society Interface 3 (11) (2006) 795–803. doi:10.1098/rsif.2006.0134.
[19] M. Courtot, W. Bug, F. Gibson, A.L. Lister, J. Malone, D. Schober, R. Brinkman, A. Ruttenberg, The OWL of biomedical investigations, in: OWLED 2008, 2008.
[20] P.L. Whetzel, H. Parkinson, H.C. Causton, L. Fan, J. Fostel, G. Fragoso, L. Game, M. Heiskanen, N. Morrison, P. Rocca-Serra, S.-A. Sansone, C. Taylor, J. White, C.J. Stoeckert, The MGED ontology: a resource for semantics-based description of microarray experiments, Bioinformatics 22 (7) (2006) 866–873.
[21] H.V. de Sompel, C. Lagoze, M. Nelson, S. Warner, R. Sanderson, P. Johnston, Adding eScience assets to the data web, in: C. Bizer, T. Heath, T. Berners-Lee, K. Idehen, Linked Data on the Web, LDOW2009, 2009.
[22] D. Shotton, Cito, the citation typing ontology, Journal of Biomedical Semantics 1 (suppl. 1) (2010) S6. doi:10.1186/2041-1480-1-S1-S6. URL: http://www.jbiomedsem.com/content/1/S1/S6.
[23] P. Missier, S.S. Sahoo, J. Zhao, A. Sheth, C. Goble, Janus: from workflows to semantic provenance and linked open data, in: Procs IPAW 2010, 2010.
[24] P. Jain, P. Hitzler, P. Yeh, K. Verma, A. Sheth, Linked data is merely more dbata, in: Linked AI: AAAI Spring Symposium "Linked Data Meets Artificial Intelligence", 2010.
[25] W3C OWL Working Group, OWL 2 web ontology language document overview, W3C Recommendation, World Wide Web Consortium, 2009. URL: http://www.w3.org/TR/owl2-overview/.
[26] D. Brickley, R. Guha, RDF vocabulary description language 1.0: RDF schema, W3C Recommendation, World Wide Web Consortium, 2004. URL: http://www.w3.org/TR/owl-guide/.
[27] D. De Roure, C. Goble, S. Aleksejevs, S. Bechhofer, J. Bhagat, D. Cruickshank, P. Fisher, D. Hull, D. Michaelides, D. Newman, R. Procter, Y. Lin, M. Poschen, Towards open science: the myExperiment approach, Concurrency and Computation: Practice and Experience 27 (17) (2010) 2335–2353. doi:10.1002/cpe.1601. URL: http://onlinelibrary.wiley.com/doi/10.1002/cpe.1601/abstract.
[28] D. De Roure, C. Goble, Lessons from myExperiment: research objects for data intensive research, in: Microsoft e-Science Workshop, Pittsburgh, US, 2009.
[29] P. Rocca-Serra, M. Brandizi, E. Maguire, N. Sklyar, C. Taylor, K. Begley, D. Field, S. Harris, W. Hide, O. Hofmann, S. Neumann, P. Sterk, W. Tong, S.-A. Sansone, ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level, Bioinformatics 26 (18) (2010) 2354–2356. doi:10.1093/bioinformatics/btq415.
[30] J.O. Dada, I. Spasić, N.W. Paton, P. Mendes, SBRML: a markup language for associating systems biology data with models, Bioinformatics 26 (7) (2010) 932–938. doi:10.1093/bioinformatics/btq069.
[31] I. Buchan, S. Sufi, S. Thew, I. Dunlop, U. Hiroeh, D. Canoy, G. Moulton, J. Ainsworth, A. Dale, S. Bechhofer, C. Goble, Obesity e-Lab: connecting social science via research objects, in: 2009 Int. Conf. on e-Social Science, Cologne, Germany, 2009.
[32] E. Gamma, R. Helm, R. Johnson, J. Vlissides, Design Patterns: Elements of Reusable Object-Oriented Software, in: Professional Computing Series, Addison-Wesley, 1995.
[33] H.V. de Sompel, R. Sanderson, M.L. Nelson, L. Balakireva, H. Shankar, S. Ainsworth, An http-based versioning mechanism for linked data, CoRR abs/1003.3661.
[34] E. Prud'hommeaux, A. Seabore, SPARQL Query Language for RDF, W3C recommendation, World Wide Web Consortium, 2008. URL: http://www.w3.org/TR/rdf-sparql-query/.
[35] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, J. Morissette, Bio2rdf: towards a mashup to build bioinformatics knowledge systems, in: Semantic Mashup of Biomedical Data, Journal of Biomedical Informatics 41 (5) (2008) 706–716. doi:10.1016/j.jbi.2008.03.004. URL: http://www.sciencedirect.com/science/article/B6WHD-4S3S2HJ-2/2/e60cdd078bdf86af287bc5926bcd1254.
[36] L. Sauermann, R. Cyganiak, Cool URIs for the semantic web, W3C interest group note, World Wide Web Consortium, 2008. URL: http://www.w3.org/TR/cooluris/.
[37] C. Lagoze, H.V. de Sompel, P. Johnston, M. Nelson, R. Sanderson, S. Warner, ORE specification—abstract data model, Tech. Rep., Open Archives Initiative, 2008. URL: http://www.openarchives.org/ore/datamodel.
[38] C. Lagoze, H.V. de Sompel, ORE user guide—resource map implementation in RDF/XML, Tech. Rep., Open Archives Initiative, 2008. URL: http://www.openarchives.org/ore/rdfxml.
[39] D. Newman, S. Bechhofer, D. De Roure, myExperiment: an ontology for e-research, in: Sem Web Apps in Scientific Discourse, W/Shop at ISWC 2009, 2009.
[40] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. Stephan, J.V. den Bussche, The open provenance model core specification (v1.1), Future Generation Computer Systems 27 (6) (2011) 743–756. doi:10.1016/j.future.2010.07.005. URL: http://www.sciencedirect.com/science/article/B6V06-50J9GPP-3/2/09d841ac888ed813ccc3cce84383ce27.
[41] K. Cheung, J. Hunter, A. Lashtabeg, J. Drennan, SCOPE: a scientific compound object publishing and editing system, International Journal of Digital Curation 3 (2) (2008).
[42] M. Gamble, C. Goble, Standing on the shoulders of the trusted web: trust, scholarship and linked data, in: Proceedings of the Web Science Conference 2010 WebSci10 Extending the Frontiers of Society, 2010. URL: http://journal.webscience.org/312/2/websci10_submission_72.pdf.
[43] J. Golbeck, A. Mannes, Using trust and provenance for content filtering on the semantic web, in: T. Finin, L. Kagal, D. Olmedilla (Eds.), WWW'06 W/Shop on Models of Trust for the Web, vol. 190, 2006. CEUR-WS.org.
[44] Y. Gil, V. Ratnakar, Trusting information sources one citizen at a time, in: ISWC'02: First Int. Semantic Web Conf., Springer-Verlag, London, UK, 2002, pp. 162–176.
[45] O. Hartig, J. Zhao, Using web data provenance for quality assessment, in: Int. W/Shop on Semantic Web and Provenance Management, Washington, DC, USA, 2009.

Sean Bechhofer is a Lecturer in the Information Management Group of the University of Manchester. His research interests cover tools and infrastructure to support the use of knowledge representation languages. He has developed applications, editors, parsers, APIs and interfaces to support the use of semantic technologies and participated in standardisation activities for W3C, contributing to both OWL and SKOS.

Iain Buchan is Professor of Public Health Informatics at the University of Manchester, Director of the Northwest Institute for Bio-Health Informatics, Chief Scientific Officer for North West e-Health, and an honorary Consultant in Public Health in the English National Health Service. He has backgrounds in clinical medicine, pharmacology, public health and computational statistics, and runs a multi-disciplinary team bridging health sciences, computer science, statistics, social science, and management science. His work centres on harnessing routinely-collected health data and building usefully-complex models for developing care-services, improving clinical research and providing public health intelligence. He also drives informatics innovation to support personal and co-produced healthcare decision making.

Please cite this article in press as: S. Bechhofer, et al., Why linked data is not enough for scientists, Future Generation Computer Systems (2011), doi:10.1016/j.future.2011.08.004

Dave De Roure is a Professor of e-Research in the Oxford e-Research Centre and UK National Strategic Director for Digital Social Research. His research projects draw on Web 2.0, Semantic Web, workflow and pervasive computing technologies and he focuses on the co-evolution of digital technologies and research methods in and between multiple disciplines.

Paolo Missier is a Lecturer, now based at the University of Newcastle. His background is in data and information management (including the semantic variety), and his current research interests include the modelling and design of data and systems architectures in support of computational science. His favourite technology areas include semantic data modelling (RDF), data mining, distributed SW architectures and cloud computing.

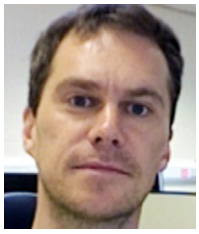
John Ainsworth is a Research Fellow in the School of Community Based Medicine, specialising in the application of emerging computing technologies to a wide range of health care challenges from predictive modelling of population needs to novel therapeutic interventions.

Jiten Bhagat is a core developer at the University of Manchester on the myExperiment and BioCatalogue projects.

Philip Couch is a software engineer, developing health information systems for the CLAHRC project. In this role, Philip forms part of a collaboration between the University of Manchester and the NHS, applying state of the art technologies to the analysis of NHS care pathways, including the measurement of health care inequalities. Philip is involved in the design and implementation of software that can be used to simulate the impacts of making changes to care pathways with the aim of significantly improving the health of the UK population.

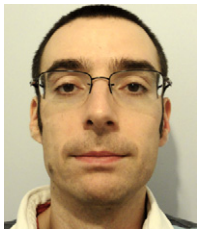
Don Cruickshank is a senior research fellow in the Web and Internet Science group at the University of Southampton. He develops software for the myExperiment social networking site for scientists.

Mark Delderfield is a Technical Project Manager for the Northwest Institute for BioHealth Informatics (NIBHI). He is responsible for leading a team to develop an e-Infrastructure that provides informatics support for clinical, medical and biological researchers.



Ian Dunlop is a software Engineer at the University of Manchester working on the Obesity e-Lab project and is the lead designer of MethodBox.

Matthew Gamble is a Ph.D. student at the University of Manchester. He is currently involved in the e-Laboratories initiative helping to define Research Objects and is particularly focused on the issues of Trust (reputation, content and provenance based trust) in eScience, web-based scientific collaboration (Science 2.0), and the Semantic Web.



Danius Michaelides is a Senior Research Fellow within the Web and Internet Science group at the University of Southampton where he is involved in building e-Science applications. His main interests are distributed systems and open information systems.

Stuart Owen is a principle developer at the University of Manchester. He is the lead designer of Sysmo-DB, a systems biology project using the Ruby on Rails framework.

David Newman is a researcher in the Intelligence, Agents and Multimedia research group at the University of Southampton. He is currently working on the NeuroHub project and has research interests including e-Science/e-Research, Social Networking, Semantic Web Technologies and Question-Answering Systems.

Shoaib Sufi is a Project Portfolio Manager at the University of Manchester. He is currently responsible for the Technical Software Project Management across a number of projects including the MethodBox system developed as part of the UK ESRC funded Obesity e-Lab project.

Carole Goble is a full professor in the University of Manchester School of Computer Science, where she co-leads the Information Management Group. She has worked closely with life scientists for many years and has an international reputation in the Semantic Web, e-Science and Grid communities. Carole is the Director of the myGrid project, a team that produce and use a suite of tools designed to "help e-Scientists get on with science and get on with scientists".

Please cite this article in press as: S. Bechhofer, et al., Why linked data is not enough for scientists, Future Generation Computer Systems (2011), doi:10.1016/j.future.2011.08.004