Research Objects: Towards Exchange and Reuse of Digital Knowledge

Sean Bechhofer[1], David De Roure[2], Matthew Gamble[1], Carole Goble[1], Iain Buchan[1]
[1]University of Manchester
[2]University of Southampton
sean.bechhofer@manchester.ac.uk

ABSTRACT

What will researchers be publishing in the future? Whilst there is little question that the Web will be the publication platform, as scholars move away from paper towards digital content, there is a need for mechanisms that support the production of self-contained units of knowledge and facilitate the publication, sharing and reuse of such entities. In this paper we discuss the notion of *research objects*, semantically rich aggregations of resources, that possess some scientific intent or support some research objective. We present a number of principles that we expect such objects and their associated services to follow.

1. INTRODUCTION

Changes are occurring in the ways in which scientific research is conducted. Within e-laboratories, 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 does not, however, necessarily make a scientific output decomposable. Nor does it guarantee that outputs, results or methods are *reusable*.

This is exemplified as follows – 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.

Studies continue to show that research in all fields is increasingly collaborative [12]. 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 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.

A recent illustrative example of the value of open "data publication" can be seen in the furore surrounding freedom of information and evidence in climate change research[1]. The UK Information Commissioner's Office decided that, by refusing to comply with requests for data concerning claims by its scientists that man-made emissions were causing global warming, a university research unit breached the Freedom of Information Act. Setting aside the details of this particular case, encouraging the principled publication of data with the analysis and conclusions would have helped avoid the "data silos" that led to the necessity of a Freedom of Information request in the first place. Greater transparency for the basis and veracity of the climate modelling would have been achieved without removing the intellectual property and academic rewards for doing the modelling.

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.

If not traditional papers and volumes, what, then, *should* researchers be publishing? Whilst the digital exchange of data is straightforward, the digital exchange and transfer of scientific knowledge in collaborative environments has proven to be a non-trivial task [2], requiring tacit, and rapidly changing expert knowledge – much of which is lost in traditional methods of publication and information exchange. We believe that there is a need for mechanisms that support the production of self-contained units of knowledge and that facilitate the publication, sharing and reuse of such entities.

In this paper, we briefly discuss the notion of *Research Objects*, semantically rich aggregations of resources that provide the "units of knowledge" as introduced above. A Research Object (RO) provides a container for a principled aggregation of resources, produced and consumed by common services and shareable *within* and *across* organisational

WWW2010, April 26-30, 2010, Raleigh, North Carolina.

[1]BBC News Report *Climate e-mails row university 'breached data laws'* `http://news.bbc.co.uk/1/hi/uk/8484385.stm`

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. An association with a dataset (or service, or result collection, or instrument) is now more than just a citation or reference to that dataset (or service or result collection). The association is rather a *link* to that dataset (or service or result collection) that can be explicitly followed or dereferenced providing access to the actual resource and thus enactment of the service, query or retrieval of data, and so on. In addition an RO includes additional semantic information that will *organize* not just *aggregate* the resources. Note that this paper does not propose a complete technical solution. Whilst we are aware that there are many aspects of our approach that require more detailed description, we are also aware that there a number of nascent Research Objects emerging in e-laboratory solutions. This paper is therefore intended as a position paper or *manifesto*, outlining the principles and features of our approach.

2. CONTEXT

Although there is ongoing debate into the motivation and willingness of scientists to share their experimental work and data, both pre- and post-publication [10], recent application of social networking techniques to the development of e-laboratories has shown that scientists are increasingly prepared to share their experimental data and their resources [7] and in turn discover and reuse resources that have been shared by other scientists.

This work has been motivated by a key observation from a number of projects building e-laboratories to share and consume scientific resources, namely that in practice, scientific investigations comprise of collections of resources. These projects include the myExperiment Virtual Research Environment[2], SysMO SEEK[3], Obesity e-Lab and its technology platform MethodBox[4], the Greater Manchester Collaboration for Leadership in Applied Health Research and Care: Systems [5] and the National e-Infrastructure for Social Simulation (NeISS) project[6]. As a motivating example we consider the use of "Packs" within one of those projects, myExperiment. The myExperiment Virtual Research Environment [7] allows scientists to share digital items associated with their research. It provides a social web site (built using Web 2.0 principles) where scientists can discover, publish and curate scientific workflows and other artefacts. The project focused initially on workflows[7], and now embraces several workflow systems including Taverna[8], Trident[9] and Triana[10].

There was a recognition [6] that workflows can be enriched through a bundling of the workflow with additional information (e.g. input data, results or logs of workflow executions, publications). In myExperiment this is supported through

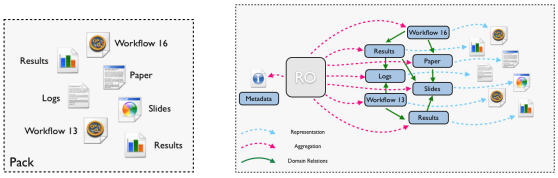

Figure 1: Resources in myExperiment pack and RO

the notion of "Packs", collections of items that can be shared as a single entity. The position is illustrated by the left of Figure 1 which shows a number of resources aggregated in a single myExperiment pack.

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 shown in the right in Figure 1, with the aggregation having been 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.

This *enrichment* of the aggregation, and the corresponding added value in terms of reuse and sharing is what we intend to achieve through the definition of Research Objects. Note that in this particular example relationships are expressed between the component entities as resources (rather than their concrete representations).

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 [4]). Ontologies such as EXPO[11], OBI [5], MGED[12], SWAN/SIOC[13] provide vocabularies that allow the description of studies and the resources that are used within them. Semantically Annotated LaTeX (SALT)[14] is a semantic authoring framework targeted at enriching scientific publications with semantic annotations. SALT defines ontologies for externalizing the rhetorical and argumentation structures captured within a publication's content. The HyPER community[15] are focused on infrastructure to support Hypotheses, Evidence and Relationships. 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 *content* rather than *container*. OAI's Object Reuse and Exchange (OAI-ORE) [9] goes some way towards providing a basic vocabulary for the description of aggregations (and is used in myExperiment's RDF export of packs [11]).

What is missing are principles and mechanisms for the description of the aggregation of resources and, through sufficient description of the contribution of these resources to the investigation and their relationships to each other, captures the additional value of the collection and enables reuse through the exchange of a single object: the *Research Object*.

In practice, during the life cycle of an investigation (which

[2] http://www.myexperiment.org
[3] http://www.sysmo.net
[4] http://www.methodbox.org
[5] http://www.healthimpact.org.uk
[6] http://www.neiss.org.uk
[7] although other content types can be shared in myExperiment
[8] http://taverna.sourceforge.net
[9] http://connect.microsoft.com/Trident
[10] http://www.trianacode.org
[11] http://expo.sourceforge.net/
[12] http://mged.sourceforge.net/ontologies
[13] http://www.w3.org/TR/hcls-swansioc/
[14] http://salt.semanticauthoring.org/
[15] http://hyp-er.wik.is/

spans activities including planning, execution of experiments, analysis of data and dissemination/publication), scientists will work with multiple content types with data distributed in multiple locations. Although potentially useful individually, when considered collectively these resources enrich and support each other and constitute a scientific investigation. 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 Research Objects. Content could include **Questions:** The question context including a description of the problem, a digest of preceding research, and optionally a hypothesis; **Organisational context:** Information about ethical approval, governance policies, the investigators involved in the experiment; **Study Design:** scientific workflows, web services, scripts; **Data:** sources and collections of cleaned or raw data in various formats ranging from flat files from signal transducers to spreadsheets and databases; **Results:** spreadsheets, SBRML Methods; **Answers:** Publications, papers, reports, slide-decks, DOIs, PUBMED ids.

3. RO PRINCIPLES & FEATURES

The goal of Research Objects is to create a class of artefacts that can encapsulate our digital knowledge and provide a mechanism for sharing and discovering assets of *reuseable* research and scientific knowledge.

We identify below a number of principles that we expect Research Objects to follow. These principles are mixed in nature and include *behaviours* that we expect Research Objects to exhibit along with *functionalities* that we expect Research Objects (and their associated services to support). The principles inform both the features of the research object model and the services that will produce, consume and manipulate research objects.

Reuse can come in many different forms. Objects can be reused as they are, they can be decomposed and then recomposed in slightly different ways. If they encapsulate processes, these processes can be re-enacted or previous executions of the process can be examined. Below, we introduce a number of principles, which are 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.

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 study or Research Object. We refer to the Research Object being reused as a whole or single entity in this case.

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 informs how they can be re-used.

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. This may involve access to data or execution of services, thus introducing a requirement for enactment services or infrastructure. In addition, the user will need sufficient privileges to access any data or services required.

Reproducible To reproduce (or replicate) a result can be for a third party to start with the same materials and methods and see if a prior result can be confirmed. It can also be for the original investigator to recreate a study prior to repurposing it. 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.

Replayable If studies are automated they might involve single investigations that happen in milliseconds or long running processes that take months. Either way, the ability to replay the study, and to examine parts of it, is essential for human understanding of what happened. Replay thus allows the investigator 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.

Traceability 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 an experiment in order to be convinced of the validity of results. Audit is required not just for regulatory purposes, but allows for the results of experiments to be interpreted and reused, thus a Research Object should provide sufficient information to support audit of the aggregation as a whole, its constituent parts, and any process that it may encapsulate.

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

Aggregation Research Objects are aggregations of content. Thus a Research Object framework needs to provide a mechanism for this aggregation. Aggregations are likely to include references to resources but there may also, however, be situations where, for reasons of efficiency or in order to support persistence, Research Objects should also be able to aggregate literal data as well as references to data.

Identity Like other information management frameworks, Research Objects require the ability to uniquely refer to an object instance or record by an identifier that is guaranteed to be unique throughout the system in which it is used. 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 or substitutable? 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 Research Object framework is grounded on the provision of machine readable and processable metadata. Research Objects will be annotated as individual objects, while metadata will also be used to describe the internal structures and relationships contained within a Research Object. 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.

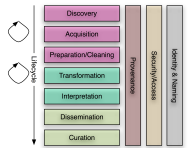

Figure 2: RO Lifecycle

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 an investigation passes. Research Objects have state, and this state may impact on available operations. For example, an study may go through a number of stages including ethical approval, data collection, data cleaning, data analysis, peer review and publication (see Figure 2). At each stage in the process, it may be possible to perform different actions on the object. Thus a principled description of Research Object lifecycle is needed. If ROs are to be used to support publication, such a lifecycle will need to take account of the *curation* of ROs, once published. **Versioning** In tandem with the question of lifecycle comes the issue of Versioning. Research Objects are dynamic in that their contents can change and be changed – additional contents may be added to aggregations, or additional metadata can be asserted about the contents or relationships between content. The resources that are aggregated may change. 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 historical points in their lifecycle.

Management Management of Research Objects requires Create, Retrieve, Update, Delete (CRUD) operations, for the creation, manipulation of those objects. Storage and indexing for discovery are also considerations.

Security Research Objects are seen as a mechanism to facilitate sharing of experiments, data and methods. With sharing come issues of access, authentication, accounting and trust that we can loosely classify as being relevant to Security.

Attribution A clear mechanism for identifying the attribution and provenance of information contained within a Research Object is necessary. Including such information will help to support credit and reward based on publication of data and methods (as opposed to papers as is currently largely the case). The question of credit and attribution also highlights the necessity that Research Objects are, in general, not just about data and methods, but also contain (links or references to) *people*.

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 Research Objects, particularly serendipitous or unpredicted reuse: the notion of "graceful degradation of understanding", whereby Research Object services are able to consume Research Objects without necessarily understanding or processing all of their internal structure or content. This places a requirement of principled extensibility on the research object model. In addition, concerns should be clearly separated so that applications or services can choose to ignore those aspects which are irrelevant. This also introduces a need for a clear characterization and differentiation between Research Object *Services* and Research *Objects*.

4. STEREOTYPES

An examination of our projects involved in e-lab related activities has allowed the identification of a number of "stereotypical Research Objects" – common patterns of resource aggregation. Below, for a selection of these, we highlight the features and principles that are important.

Publication Object One key motivation for our Research Object notion, as set out in the introduction is for objects that allow us to move from traditional paper based (linear) dissemination mechanisms, and support "rich publication". This is not simply about making works available in digital formats (e.g. online PDFs of papers), although electronic publication of course forms a piece of the activity. Rather, this is about providing aggregations that explicitly bring together the presentation of a piece of work – the "paper" – along with the evidence for the conclusions that are being presented, e.g. data sets, experimental results, the workflows used to produce those results and so on.

Publication Objects are intended as a record of activity, and should thus be *immutable*. That is not to say that versions of a Publication Object cannot be produced, but such versions should be considered distinct objects. This relates to the notion of lifecycle, with clearly defined *publication* events needed. Publication Objects must be *citeable*. As already discussed, mechanisms for identification and reference to Research Objects are required, but for publication these should be of a form which is externally usable.

Credit and *attribution* are central aspects of the publication process as they are key to providing rewards, and thus incentives, for scientific publication.

The Publication Object will also make use of ontologies for the representation of the rhetorical or argumentation structure in the publication (see Section 2).

Work Object We have used the term Work Object synonymously with Research Object where the application is beyond research, for example to business intelligence or audit - where repeatability, replayability and repurposing are key aspects[1].

Live Object Live Objects 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. Live objects are potentially under the control of multiple owners and may fall under mixed stewardship. There are thus issues relating to *security*, and *access* control.

Exposing Object Research Objects can provide a wrapper for existing data, providing a standardised metadata container. For example, within SysMO, there is widespread usage of spreadsheets to record data from an experiment. Theses spreadsheets may be gathered together and aggregated along with the methods used to produce them. This aggregation can be seen as a Research Object (including data, methods etc), but it can also be considered to be comprised of smaller, component Research Objects which wrap each spreadsheet. The Exposing Object provides a Wrapper [8] that allows the spreadsheet to be seen as a Research Object, facilitating its exposure and integration into the Web of Linked Data.

View/Context Object A View or Context object can provide a view over some already exposed data. It is here that Research Objects can interact with data that is exposed or published using Linked Data principles [3], providing a "Named Graph" for those resources.

Method Object A Method Object reports methodolog-

ical research in a Research Object and exposes the method for easy consumption by other Research/Work Objects. This may be a key feature for propagating methodological integrity and avoiding translation errors for methods.

Archived Object An Archived Object encapsulates an aggregation that is in some way "finished", deprecated or no longer "live". Archived Objects should thus be *immutable*, with no further changes allowed. For example, an Archived Object may be used to collect together and record resources used in an experiment which has been abandoned. Archived Objects are similar to Publication Objects, but may not require the same level of detail in terms of, for example credit and attribution.

5. CONCLUSIONS

Traditional paper publication – or even electronic publication following the "paper metaphor" – will not adequately support reusuable, shared research. New mechanisms are needed that will allow us to share, exchange and reuse digital knowledge as (de)composable entities. Our solution to this is *Research Objects*, semantically rich aggregations of resources that bring together the data, methods and people involved in (scientific) investigations.

Research Objects will allow scientists to group together and associate the resources used in their work. This will then lead to greater *transparency*, allowing for the *validation* of results. Ideally, this should also lead to greater *sharing* of resources and the *reuse* of existing data sets and methods.

We have approached the topic from two different directions. A number of existing projects are already beginning to apply a Research Objects approach to the organisation and publication of their data. At the same time, we have been reflecting on how such aggregations might play a part in the scientific process, which leads us to the principles as identified in Section 3. Our Research Object view provides a layer of aggregation structure that sits well with the Linked Data view of the world, as advocated within the Semantic Web community[3]. ROs are both themselves resources accessible via linked data principles, and will aggregate linked data resources.

We have given a brief overview of our motivation and position here – there are, of course, open questions and issues.

Credit, attribution and rewards A key aspect of myExperiment is its credit and attribution model. Included at the request of domain scientists, it allows for credit to be made for derivative works. A shift to RO based publishing would require a similar re-engineering of reward structures for scientists – citation counts are no longer enough, if works are also built on reuse or repurposing of data and methods.

Trustworthiness and Quality Trust is a challenge common to all emerging collaborative environments that promote open science and the rapid exchange of experimental and pre-publication data and methods. How can consumers trust the user generated content and how can producers of content trust users to consume, interpret and attribute correctly? As an identifiable container, Research Objects allow us to compute and attribute measure of trust to the object itself, with potential to apply and extend methods for modeling and computing social trust, trust in content and trust based on provenance information.

In closing, we believe that the Research Objects approach will allow us to conduct scientific research in ways that are *efficient*, in that it costs less to borrow a model than to recreate it; *effective*, supporting larger scale research by reusing parts of models; and *ethical*, as research supported by public funds will provide benefits not just for individual scientists but for a wider community.

Acknowledgments This work has been influenced by discussions within the e-laboratories technical architecture group (e-lab TAG)[16]

6. REFERENCES

[1] J. Ainsworth and I. Buchan. e-Labs and Work Objects: Towards Digital Health Economies. In *Communications Infrastructure. Systems and Applications in Europe.*, volume 16 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering.*, pages 206–216. Springer, 2009.

[2] N. Bos, A. Zimmerman, J. Olson, J. Yew, J. Yerkie, E. Dahl, and G. Olsen. From shared databases to communities of practice: A taxonomy of collaboratories. *Journal of Computer-Mediated Communication*, 12(2), 2007.

[3] T. H. Christian Bizer and T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, Special Issue on Linked Data(*in press*), 2009.

[4] T. Clark, J. S. Luciano, M. S. Marshall, E. Prud'hommeaux, and S. Stephens, editors. *Semantic Web Applications in Scientific Discourse 2009*, volume 523. CUER Workshop Proceedings, October 2009.

[5] M. Courtot, W. Bug, F. Gibson, A. L. Lister, J. Malone, D. Schober, R. Brinkman, and A. Ruttenberg. The OWL of Biomedical Investigations. In *OWLED 2008*, 2008.

[6] D. De Roure and C. Goble. Lessons from myexperiment: Research objects for data intensive research. In *Microsoft e-Science workshop*, 2009.

[7] D. De Roure, C. Goble, and R. Stevens. The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows. *Future Generation Computer Systems*, 25:561–567, 2009.

[8] E. Gamma, R. Helm, R. Johnson, and J. Vlissides. *Design Patterns: Elements of Reusable Object-Oriented Software*. Professional Computing Series. Addison-Wesley, 1995.

[9] C. Lagoze, H. V. de Sompel, P. Johnston, M. Nelson, R. Sanderson, and S. Warner. ORE Specification - Abstract Data Model. Technical report, Open Archives Initiative, 2008.

[10] B. Nelson. Data sharing: Empty archives. *Nature*, 461(7261):160–3, 2009.

[11] D. Newman, S. Bechhofer, and D. De Roure. myExperiment: An ontology for e-Research. In *Semantic Web Applications in Scientific Discourse, Workshop at ISWC 2009*, 2009.

[12] G. Olson, A. Zimmerman, and N. Bos. *Scientific Collaboration on the Internet*. MIT Press, 2008.

[16]The e-lab TAG consists of members of the Universities of Manchester and Southampton who are currently involved in the development and realisation of e-laboratories. Those who contributed include John Ainsworth, Jiten Bhagat, Phillip Couch, Don Cruickshank, Mark Delderfield, Ian Dunlop, Danius Michaelides, Paolo Missier, Stuart Owen, David Newman, and Shoaib Sufi.