2010 Sixth IEEE International Conference on e–Science

Why Linked Data is Not Enough for Scientists

Sean Bechhofer[1], John Ainsworth[2], Jiten Bhagat[1], Iain Buchan[2], Philip Couch[2],
Don Cruickshank[3], David De Roure[3,4], Mark Delderfield[2], Ian Dunlop[1],
Matthew Gamble[1], Carole Goble[1], Danius Michaelides[3],
Paolo Missier[1], Stuart Owen[1], David Newman[3], Shoaib Sufi[1]

[1]School of Computer Science, University of Manchester, UK
[2]School of Medicine, University of Manchester, UK
[3]School of Electronics and Computer Science, University of Southampton, UK
[4]Oxford e-Research Centre, University of Oxford, UK

***Abstract*—Scientific data stands to represent a significant portion of the linked open data cloud and science itself stands to benefit from the data fusion capability that this will afford. However, simply publishing linked data into the cloud does not necessarily meet the requirements of reuse. Publishing has requirements of provenance, quality, credit, attribution, methods in order to provide the *reproducibility* that allows 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.**

I. Introduction

Changes are occurring in the ways in which scientific 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 does not, however, 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. The models used to support medical decisions are not dynamically linked to the body of knowledge that defines best practice. More than half of the effects of medical treatments cannot be predicted from the literature, because trials exclude women of child bearing age, people with other diseases or on other medications. Doctors audit the outcomes of their treatments using research methods yet the results are not captured and put back into medical research for the benefit of society [1].

As an 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 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 respect research methodology nor would it respect the flow of 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, semantically rich aggregations of resources that provide the "units of knowledge" which supply structure for delivery of information as Linked Data. 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 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 case studies 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

978-0-7695-4290-4/10 $26.00 © 2010 IEEE
DOI 10.1109/eScience.2010.21

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.

II. Linking knowledge, 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.

Mesirov [4] describes the notion of Accessible Reproducable Research, where scientific publications should provide clear enough descriptions of the protocols to enable allow 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" publication can present intellectual arguments, fostering reinforcement requires inclusion of data, methods and results in our publications, thus supporting reproducibility. A problem with traditional paper publication, 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, techniques such as text mining are used to extract the knowledge from the papers, resulting in a loss of that knowledge.

The 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) which can be followed or dereferenced explicitly, thereby providing access to the actual resource and thus enactment of the service, query or retrieval of data, and so on.

Linked Data As outlined above, providing links, rather than associations, between resources will help foster 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 VII.

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 [7]). Ontologies such as EXPO [8], OBI [9], MGED [10], SWAN/SIOC [11] provide vocabularies that allow the description of experiments and the resources that are used within them. The HyPER community 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*.

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

Alice runs an (in-silico) experiment that involves the execution of a scientific workflow over some data sets. The output of the workflow includes results of the analysis along with provenance information detailing the services used, intermediate results, logs and final results. She collects together and publishes this information as a Research Object 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 or LarKC, providing additional context. Alice embeds the RO in a blog post.

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.

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

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, [12] 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 Research Objects including aspects such as lifecycle, ownership, versioning, etc. It thus says little about how that

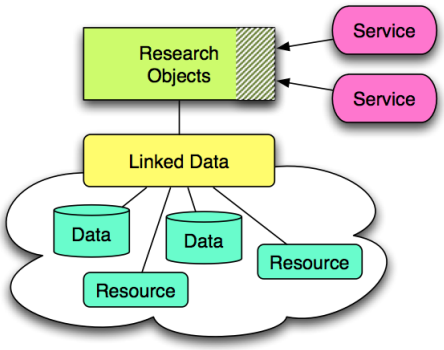

Fig. 1. Research Object Layer

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 [13] 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. 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 Figure 1 with the Research Object Layer providing a structured "view" on the underyling resources that can then be consumed by RO aware services.

What is missing, then is a mechanism that describes 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. Our notion of *Research Objects* are intended to supply these aggregations and provide a container infrastructure, facilitating the sharing and reuse of scientific data and results. Such a common model then facilitates the construction of services for the creating, manipulation and sharing of our research results.

III. Characterising Reuse

In our scenario, we assert that Bob wants to reuse Alice's results and observe that the term "re-use" can be used to describe a range of activities, and 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 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. As introduced above, *reproducability* is key in supporting the validation of experiments or procedures.

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 ROs must be reusable as part of a new experiment or RO. By reuse here, we refer to a "black box" consideration of the RO where it is to be reused as a whole or single entity.

Repurposeable Reuse of an RO may also involve the reuse of constituent parts of the RO, 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 is a clue to how they can be re-used. In order to allow such a "dis-agreggation" and recombination, ROs should expose their constituent pieces. Thus our RO framework also has need of an aggregation mechanism.

Repeatable There should be sufficient information in an RO 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. Repeat may involve access to data or execution of services, thus introducing a requirement for enactment services or infrastructure that can consume ROs. 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 materials 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 long running 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 must be referenceable or citeable. Thus mechanisms are needed that allow us to refer unambiguously 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 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 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.

IV. RO Principles, Behaviours and 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. As discussed above, ROs are intended to support reuse in a number of ways. These various kinds of reusability 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 inform the features of the research object model and the services that will produce, consume and manipulate research objects.

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 should allow for references 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 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. 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 a 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.

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.

Management Management of ROs requires Create, Retrieve, Update, Delete (CRUD) operations, for the creation, manipulation of those objects. Storage is also a consideration.

Security ROs are seen as a mechanism to facilitate sharing of experiments, data and methods. 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 Research Object model.

V. 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. Scientists utilise 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 [14]. Contents might include

Questions A research problem, a hypothesis;

Data Data Sets; measurements, database records, spreadsheets;

Results Spreadsheets, SBRML Methods;

Experimental Design Scientific workflows, scripts;

Organisational Context Ethical approval; governance policies; investigators;

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

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.

myExperiment The myExperiment Virtual Research Environment 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 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 the course of 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 [15] 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 view point – 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 VII.

SysMO SEEK Systems Biology of Microorganisms (SysMO)[1] is a European trans-national research initiative, consisting of 91 institutes organized into eleven projects whose goal is to create computerized mathematical models of the dymanic molecular processes occurring in microorganisms. SysMO-DB[2] 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[3] is an "assets catalogue" describing data, models, Standard Operating Procedures (SOPs), workflows and experiment descriptions. Yellow Pages provide directories of the people who are involved in the project. The JERM (Just Enough Results Model) allows the exchange, interpretation and comparison of different types of data and results files across SysMO. 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 specializes, generic components taken from the myExperiment project.

A number of challenges characterize 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. 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 constuct and test the model, within the context of one or more Assays. 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.

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. With greater use of workflows in the future, *repeatability* and *replayability* will begin to play a part.

[1] http://www.sysmo.net/

[2] http://www.sysmo-db.org/

[3] http://www.sysmo-db.org/seek

Methodbox Methodbox is a generic solution for cross disciplinary survey based research arising from the Obesity e-Lab project [16]. This project is focused on improving the understanding of obesity between social scientists, health scientists and public health professionals, thereby supporting better research and policy decisions. Traditionally social and health scientists have shared common sources of data such as the Health Surveys for England via the UK Data Archive, but have worked largely separately. Public health policy, however, requires a hybrid social and health perspective on major problems such as obesity. Thus Methodbox was developed as a technology platform to turn "data archives into data playgrounds", thereby encouraging collaboration across the disciplines that use the archives. 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.

In Methodbox ROs are invisible, but they are intended for import and export between Methodbox and other e-Laboratories such as the NHS e-Lab (www.newh.org.uk) that public health professionals are using for sharing data, methods and expertise behind the NHS firewall. Thus academia can bolster the methodological expertise in the public health service by sharing ROs along a service bus. Attribution, sharing and audit logs will become particularly important for cross organisation as well as cross discipline sharing. So Methodbox is taking the "RO on the inside" approach, anticipating future value of reuse and audit of the semantic aggregation of research entities.

VI. Stereotypes

An examination of our projects involved in e-Laboratory related activities has allowed the identification of a number of "stereotypical Research Objects" – common patterns of resource aggregation.

Publication Objects One key motivation for our RO 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 but is rather 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. Publication Objects are intended as a record of activity, and should thus be immutable with versions be considered as distinct objects. This relates to the notion of lifecycle, with clearly defined publication events needed. Publication Objects must be citeable. 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 II).

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

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 Objects ROs 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. These spreadsheets may be gathered together and aggregated along with the methods used to produce them. This aggregation can be seen as a RO (including data, methods etc), but it can also be considered to be comprised of smaller, component ROs which wrap each spreadsheet. The Exposing Object provides a Wrapperthat allows the spreadsheet to be seen as a 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 report methodological research in a RO and expose 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 Objects encapsulate 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.

We are already observing the use of myExperiment packs as Publication Objects, for example collecting Workflows along with results obtained or papers along with presentational materials. Packs also serve as Archive Objects. SysMO provides Exposing Objects wrapping spreadsheet data, while MethodBox is allowing the sharing of Method Objects (scripts).

The OAIS model [17] also identifies variants of aggregation such as *dissemination* and *archival* information packages, corresponding loosely to our notion of publication or archived objects.

VII. Implementing ROs: Linked Data and OAI-ORE

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:

1) Use URIs as names for things
2) Use HTTP URIs so that people can look up those names.

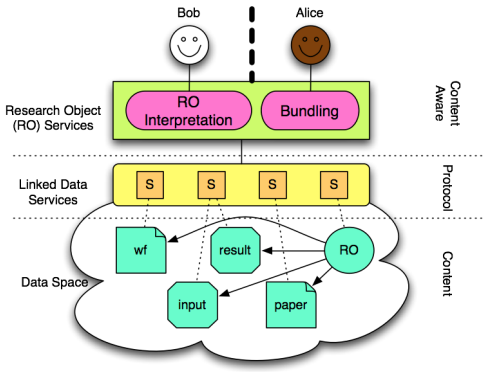

Fig. 2. Detailed Layers

3) When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
4) Include links to other URIs. so that they can discover more things.

Research Objects 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. In particular, the separation of the identity of an RO from serializations of the description of its content reflects the handling on non-information resources – we consider a particular RO to be a non-information resource which may have alternative concrete representations.

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 [18]). ORE defines a data model and a number of concrete serializations (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 aggregrations of 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 [19], which integrates well with current approaches towards the publication of Linked Data [11]. Our latest work in myExperiment makes use of the OAI-ORE vocabulary and model in order to deliver ROs in a Linked Data friendly way [20]. Although specific to myExperiment, the following discussion is pertinent to the other e-Laboratories.

Packs are created using a shopping basket (or wishlist) 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[4], entities in myExperiment are considered as Non-Information Resources and 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[5]). In addition to this "Linked Data" publishing, myExperiment content is also available through a SPARQL endpoint[6] 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 – instead of asking us to provide this in the API. 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 our 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 healthcare and life sciences area [14]. 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.

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. A RO is essentially an aggregation of resources, and we are using ORE as the basis for describing our ROs. Specific vocabulary can be defined which extends the ORE relationships and is then used to describe the relationships between the aggregated resources.

The ROs Upper Model (ROUM) provides basic vocabulary that is used to describe general properties of Research Objects that can be shared across generic e-Laboratory services. For example, the basic lifecycle states of ROs (as described in 3.1) are described in this upper model. RO 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

[4]http://www.w3.org/TR/cooluris/
[5]http://openprovenance.org/
[6]http://rdf.myexperiment.org

behaviour. Applications which are not aware of the vocabulary used may still, however be able to operate on the overall aggregation structure. This layered approach then helps meet our principle for graceful degradation of understanding across e-Laboratory services (see Section IV).

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. 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. Figure 2 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.

VIII. Discussion

A number of open questions and issues require further investigation

Credit, attribution and rewards. A key aspect of e-Laboratores is user-visibility, credit and attribution. Included at the request of domain scientists, this model 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. 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, ROs allow us to attribute a measure of trust to the object itself, with potential to apply and extend methods for modeling and computing social trust [21], trust in content [22] and trust based on provenance information [23].

Encapsulation and Versioning. When a scientist returns to an RO, it must refer to the same versions of the data that were originally used. Neither Linked Open data nor OAI-ORE tackle versioning explicitly. We envisage appropriate tooling to support archiving and validation of ROs.

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 Research Objects, semantically rich aggregations of resources that bring together the data, methods and people involved in (scientific) investigations.

A number of existing projects are already beginning to apply an RO approach to the organisation and publication of their data. In particular, myExperiment has a notion of a prototypical RO (the pack), and the capability to export this 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 RO view provides a layer of aggregation structure that sits well with the Linked Data view of the world. ROs are both themselves resources accessible via linked data principles, and will aggregate linked data resources.

In closing, we believe that the Research Objects 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.

References

[1] J. Ainsworth and 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, and N. Bos, *Scientific Collaboration on the Internet*. MIT Press, 2008.

[3] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data - The Story So Far," *Int. J. on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1–22, 2009.

[4] J. P. Mesirov, "Accessible reproducible research," *Science*, vol. 327, no. 5964, pp. 415 – 416, January 2010.

[5] D. De Roure and C. Goble, "Anchors in Shifting Sand: the Primacy of Method in the Web of Data," in *Web Science Conference 2010, Raleigh NC*, April 2010.

[6] B. Mons, "Which gene did you mean?" *BMC Bioinformatics*, vol. 6, p. 142, 2005.

[7] T. Clark, J. S. Luciano, M. S. Marshall *et al.*, Eds., *Semantic Web Applications in Scientific Discourse 2009*, vol. 523. CUER Workshop Proceedings, October 2009.

[8] L. N. Soldatova and R. D. King, "An ontology of scientific experiments." *J. of the Royal Society, Interface / the Royal Society*, vol. 3, no. 11, pp. 795–803, December 2006.

[9] M. Courtot, W. Bug, F. Gibson *et al.*, "The OWL of Biomedical Investigations," in *OWLED 2008*, 2008.

[10] P. L. Whetzel, H. Parkinson, H. C. Causton *et al.*, "The MGED Ontology: a resource for semantics-based description of microarray experiments," *Bioinformatics*, vol. 22, no. 7, pp. 866–873, 2006.

[11] H. V. de Sompel, C. Lagoze, M. Nelson *et al.*, "Adding eScience Assets to the Data Web," in *Linked Data on the Web (LDOW2009)*, C. Bizer, T. Heath, T. Berners-Lee, and K. Idehen, Eds., 2009.

[12] P. Missier, S. S. Sahoo, J. Zhao *et al.*, "Janus: from Workflows to Semantic Provenance and Linked Open Data," in *Procs IPAW 2010*, 2010.

[13] P. Jain, P. Hitzler, P. Yeh *et al.*, "Linked Data is Merely More Data," in *Linked AI: AAAI Spring Symposium "Linked Data Meets Artificial Intelligence"*, 2010.

[14] D. De Roure, C. Goble, S. Aleksejevs *et al.*, "Towards open science: The myexperiment approach," *Concurrency and Computation: Practice and Experience*, 2010, *in press*.

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

[16] I. Buchan, S. Sufi, S. Thew *et al.*, "Obesity e-Lab: Connecting Social Science via Research Objects," in *2009 Int. Conf. on e-Social Science, Cologne, Germany*, 2009.

[17] Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information System (OAIS)," Open Archives Initiative, Blue Book CCDS 650.0-B-1, 2002.

[18] C. Lagoze, H. V. de Sompel, P. Johnston *et al.*, "ORE Specification - Abstract Data Model," Open Archives Initiative, Tech. Rep., 2008. [Online]. Available: http://www.openarchives.org/ore/datamodel

[19] C. Lagoze and H. V. de Sompel, "ORE User Guide - Resource Map Implementation in RDF/XML," Open Archives Initiative, Tech. Rep., 2008. [Online]. Available: http://www.openarchives.org/ore/rdfxml

[20] D. Newman, S. Bechhofer, and D. De Roure, "myExperiment: An ontology for e-Research," in *Sem Web Apps in Scientific Discourse, W/shop at ISWC 2009*, 2009.

[21] J. Golbeck and A. Mannes, "Using Trust and Provenance for Content Filtering on the Semantic Web," in *WWW'06 W/shop on Models of Trust for the Web*, T. Finin, L. Kagal, and D. Olmedilla, Eds., vol. 190. CEUR-WS.org, 2006.

[22] Y. Gil and V. Ratnakar, "Trusting Information Sources One Citizen at a Time," in *ISWC '02: First Int. Semantic Web Conf.* London, UK: Springer-Verlag, 2002, pp. 162–176.

[23] O. Hartig and J. Zhao, "Using web data provenance for quality assessment," in *Int. W/shop on Semantic Web and Provenance Management, Washington D.C., USA*, 2009.