

myExperiment: An ontology for e-Research

David R Newman¹, Sean Bechhofer², David De Roure¹

¹ School of Electronics and Computer Science, University of Southampton,
Southampton, UK

`drn05r@ecs.soton.ac.uk`

² School of Computer Science, The University of Manchester, Manchester, UK

Abstract. myExperiment describes itself as a “Social Virtual Research Environment” that provides the ability to share Research Objects (ROs) over a social infrastructure to facilitate actioning of research. The myExperiment Ontology is a logical representation of the data model used by this environment, allowing its data to be published in a standard RDF format, whilst providing a generic extensible framework that can be reused by similar projects. ROs are data structures designed to semantically enhance research publications by capturing and preserving the research method so that it can be reproduced in the future. This paper provides some motivation for an RO specification and briefly considers how existing domain-specific ontologies might be integrated. It concludes by discussing the future direction of the myExperiment Ontology and how it will best support these ROs.

1 Introduction

This paper describes the design of an OWL DL [1] ontology for myExperiment, within the context of e-Research. [2] states that myExperiment is a “Social Virtual Research Environment” (Social VRE) and as such defines the four requisite capabilities for such a system:

1. Facilitate management and sharing of *Research Objects (ROs)*
2. Support a *social model*
3. Provide an *open extensible environment*
4. Provide a platform to *action* research

Section 2 describes the myExperiment model and presents an insight to some of the motivation and design decisions taken in its construction, in particular how it has been influenced by these four capabilities. It demonstrates how this insight informed the design decisions for the ontology itself. It also explains the techniques used to try to promote reuse of the ontology in other VRE and social networking projects.

Section 3 describes the purpose of ROs in semantically enhancing research publications. It briefly considers the requirements of an RO to ensure a scientist can reproduce the research it encompasses and how its design should facilitate this. In particular how the design will provide support for integration of domain-specific ontologies / vocabularies.

myExperiment currently only supports a few basic types of RO. This paper therefore concludes by considering how the myExperiment Ontology needs to evolve to provide greater support for all ROs.

2 Building the myExperiment Ontology

2.1 The myExperiment Model

The myExperiment model has three main features that are motivated by the four capabilities of a Social VRE:

1. Content Management
2. Social Networking
3. Object Annotation

myExperiment's initial user group was bioinformaticians that wanted to be able to manage and share workflows [3]. However, it quickly became apparent that to fulfil its capabilities, a model that allows scientists from wide ranging fields to share many different types of research was needed. myExperiment allows users to manage different types of files as well as more abstract concepts and defines these items as *Contributions*.

When a user joins myExperiment, they can make friends and join *Groups*, (represented by *Friendship* and *Membership* records), to build their social network. Beyond this myExperiment allows users to send each other *Messages* and make *Announcements* to their groups. With this social network it makes it possible for Contributions to be shared in a highly customizable way through the use of additive *Policies*.

Object Annotation allows users to annotate Contributions with *Annotations*, such as *Tags*, *Ratings* *Reviews* and *Comments* to enhance search and to support curation.

2.2 Design

The purpose for building an ontology was so to produce a consistent specification for publishing myExperiment's data as RDF and contributing to the web of data. All of myExperiment's public data can be accessed via the web³ and queried using myExperiment's SPARQL endpoint⁴.

The myExperiment website has been built using the Ruby-on-Rails⁵ framework. Ruby-on-Rails was chosen because it provides a Model-View-Controller (MVC) architecture for agile development of web projects allowing rapid innovation [4]. The model component of the architecture is designed to be a thin veneer over a database engine giving the developer freedom to choose the database that suits them, in the case of myExperiment MySQL was chosen.

³ <http://rdf.myexperiment.org/>

⁴ <http://rdf.myexperiment.org/sparql>

⁵ <http://rubyonrails.org/>

Having the myExperiment model already represented in the structured form of a MySQL database provided both benefits and disadvantages when designing the myExperiment Ontology. The initial construction of the ontology was straightforward, as much of the database schema could be transcribed directly into OWL DL. A number of tools exist that can automate this process [5], (e.g. RDBtoOnto [6], DB2OWL [7], etc.). However, several factors made using such a tool unsuitable for building the myExperiment Ontology as a whole.

Ruby-on-Rails uses the database to store information to manage the web interface. In some cases this data would be inappropriate to represent, such as users' encrypted passwords and salts. In other cases this data is irrelevant such as HTTP session data.

Ruby-on-Rails supports polymorphism. This allows a Contributions table to store generic information about items users want to share and then reference tables such as Workflows and Files that store information specific to that type of Contribution. However, this means that it is impossible to determine from the database schema the different types of Contribution so that they can be represented as subclasses of Contributions within the ontology.

Capturing myExperiment's customizable sharing model was a key aspect to building the myExperiment Ontology. This is because it is one of the features that makes myExperiment unique in the VRE community. Further to this, the way that Contributions are shared greatly effects the data generated within myExperiment, such as who has tagged a Workflow or how many times a File has been downloaded. Therefore it is important that there is a way of capturing this information concisely. However, this task was complicated by having subjective groups such as friends that vary depending on the user as well as over time.

The Simple Network Access Rights Management (SNARM) Ontology⁶ allows additive policies to be defined by assigning different types of permissions to users, groups and abstract groups such as friends or all the groups a user belongs to. It is extensible to allow the definition of new types of permission and new abstract groups.

myExperiment subscribes to the Web 2.0 model of being a "perpetual beta". This means that it evolves over time as users request new features. This inevitably means the database model is not perfect. In the process of manually building the ontology it became clear where abstractions could be made and this has fed back into the design of the database model to make it simpler and more extensible.

2.3 Promoting Reuse

One of myExperiment's goals is to encourage reuse of Workflows and other Contributions. For the design of the ontology this ethos was adopted so that it both reuses existing specifications and makes itself as reusable as possible through careful consideration of design decisions.

⁶ <http://rdf.myexperiment.org/ontologies/snarm/>

A number of core ontologies / schemata already exist for representing properties and classes that exist within myExperiment. The reuse of these gives myExperiment's RDF data a graceful degradation of understanding. I.e. if a machine is presented with some myExperiment RDF, it does not have to be aware of the myExperiment Ontology to have some understanding of the type of object it is dealing with. A major task in the Semantic Web world is co-reference resolution [8]. One technique for performing this task is comparing properties, this is more likely to be successful if an object uses recognized properties from core ontologies / schemata.

The myExperiment Ontology reuses both properties and classes from Dublin Core⁷, Friend of a Friend (FOAF)⁸, Semantically Inter-linked Online Communities (SIOC)⁹ and the Open Archives Initiative's Object Reuse and Exchange (OAI-ORE)¹⁰ ontologies / schemata.

Reusing elements from core ontologies / schemata helps to promote reuse of the ontology, as it makes it easier to understand the purpose of the ontology and it gives confidence that consideration has been given to the classes and properties defined within. However, this in itself is not enough. The ontology needs to be sufficiently generic so that potential users do not disregard it for being too specific or bloated. At the same time it needs to be expressive enough to represent the whole myExperiment model. To achieve this the myExperiment Ontology was constructed as a set of modules allowing anyone who reuses them to pick and choose the modules they need. Figure 1 diagrams how the modules bolt together to build the complete myExperiment Ontology. Each module sits atop the modules it requires to define any subclass or sub-property relationships.

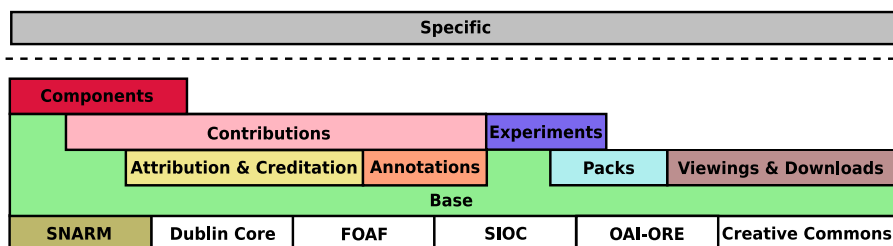


Fig. 1. myExperiment Ontology Modules Architecture

The *Base* module with the assistance of the SNARM ontology provides the bulk of the features described in section 2.1, the remaining modules support the capability to provide an *extensible environment* with definitions for specific

⁷ <http://dublincore.org/>

⁸ <http://www.foaf-project.org/>

⁹ <http://sioc-project.org/>

¹⁰ <http://www.openarchives.org/ore/>

Contributions, Annotations, usage statistics, etc. In particular the *Experiment* module is designed to represent the process of *actioning research*. The *Specific* module sits over the top and amalgamates all these modules using OWL's import property to generate an ontology for the whole myExperiment model. It also provides classes and objects that are highly specific to myExperiment instance at <http://www.myexperiment.org/>.

A more detailed document describing the ontology specification can be found at <http://rdf.myexperiment.org/ontologies/>.

3 Research Objects

3.1 Motivation

At present the accepted way of publishing research is to have a paper accepted by a conference or journal. These papers are just text documents. They may reference tools and data sources that were used and result sets that were produced. However, there is no certainty that these references will be sufficient to reproduce the research the paper describes. Even if they are, there is no guarantee that these references will still resolve to the item the paper described.

Even if it is possible to repeat the research of the paper, this has overlooked the problem of actually finding this relevant paper in the first place. [9] describes how in the field of bioinformatics there are already so many papers that it is quite difficult to find the one that discusses a particular gene in a specific context. Text-mining tools can help with this task but this begs to question why was it buried in the first place.

A Semantic Web approach can assist in such a task, by associating interoperable metadata with these papers. Much of this process can quite easily be automated but when it comes to defining the interrelationships of resources (e.g. data sources, tools and result sets), this requires both an extensible specification and a sea-change in attitudes to what constitutes a research publication.

3.2 ROs in myExperiment

Building the myExperiment Ontology has helped to clarify the concept of ROs and how myExperiment “facilitates [the] management and sharing” of them. myExperiment currently has three entities that could be considered ROs: *Workflows*, *Packs* and *Experiments*.

When a Workflow is uploaded various supporting files can also be uploaded or automatically generated, such as SVG and PNG visualizations of the Workflow, all these resources are grouped together. New *WorkflowVersions* can also be uploaded and are then associated with the original Workflow.

Packs are designed to allow users to collaboratively aggregate resources together by hand. They very closely resemble OAI-ORE Aggregations (see section 3.3). A Pack contains items that are either myExperiment Contributions or external resources. Packs allow metadata to be recorded about these items that is only relevant within the context of that Pack.

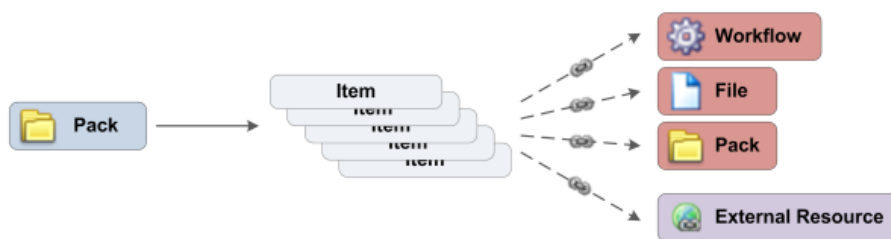


Fig. 2. The Anatomy of a Pack (credit to Jiten Bhagat)

Experiments are aggregations of *Jobs* which in themselves are aggregations of an enacted Workflow with its inputs, outputs and the *Runner* used to enact it.

These three entities currently require an explicit definition in the database structure / ontology. Supporting all conceivable ROs in this way is impractical. Rather ROs should describe their own structure so that this does not need to be defined in any system that stores them.

3.3 Anatomy of a Research Object

Research Objects (ROs) are designed to aggregate together resources, (e.g. datasets, workflows, etc.), to represent an investigation, experiment or question [10]. They need to be machine-processable so that they can be automated and demonstrate the research they encompass.

OAI-ORE is a specification defining how resources can be aggregated [11]. It was designed for the Open Repositories community to allow them to exchange objects between repositories in a standard format. OAI-ORE's first class object is an *Aggregation* that can then have one or more items associated with it as *Aggregated Resources*. Aggregations can be serialized into concrete syntax, such as RDF, Atom Feeds or RDFa, as *Resource Maps*. The specification allows for metadata to be assigned to all these objects. Metadata pertinent to Aggregated Resources only in the context of the Aggregation can also be assigned using *Proxies*.

OAI-ORE provides a suitable mechanism for making ROs machine-readable, to make them machine-processable a further specification is needed to help define the interrelationships between resources and the provenance, lifecycle, sharing, curation and usage profiles. The e-Laboratory Technical Architecture Group¹¹ are currently working on defining such a specification.

¹¹ A collaboration between VRE and related projects at the University of Manchester and University of Southampton.

Domain-specific ontologies already exist for representing entities comparable to ROs, such as the MGED Ontology¹² and the SWAN Ontology¹³. The SWAN Ontology is designed to represent scientific discourse, it has extensions for supporting domain-specific concepts such as Life Science Entities and the Gene Ontology [12]. Scientific discourse is an important component in defining reproducible research within an RO.

In many cases it should be possible to use these domain-specific ontologies or modules thereof to describe the resources and interrelationships of the RO with minimal integration required. Otherwise a simple specification of these domain-specific classes and properties could be defined to accomplish this task and integrate with the RO specification ontology.

4 Conclusion

myExperiment stores a large amount of structured data and required a way of providing this data in a standard, consistent and atomic way to make integration with similar systems easier. Building the myExperiment Ontology has facilitated delivering this data as RDF and allowing it to be queried using SPARQL. Having an ontology provides a machine-readable specification that mechanisms such as Representational State Transfer (REST) APIs do not have.

The design of the myExperiment Ontology is a continual process but through careful analysis of the underlying data model it has been designed in a generic way to ensure that future modifications should be additions or minor alterations rather than significant structural changes.

Through the reuse of existing classes and properties from core ontologies / schemata and the modularization of the ontology, a concerted effort has been made to promote reuse of the ontology in similar projects. Modularization should also allow any significant modifications to the myExperiment model to be isolated in new modules rather than requiring changes to existing ones.

In particular, as the concept of ROs becomes more evolved a new module can be constructed to support them. This module will need to integrate with the RO specification ontology to provide an interface with the provenance, lifecycle, sharing, curation and usage profiles stored within an RO, allowing the user to manage these within myExperiment.

If ROs are to be successful and provide a mechanism for publishing reproducible research there needs to be a system to allow their management and sharing prior to as well as post publication. This system should also facilitate the collaborative building of ROs, where multiple researchers, potentially situated in different locations, are involved in the same experiment or investigation. Integrating the myExperiment Ontology with the RO specification ontology should make myExperiment well placed to achieve this.

¹² <http://mged.sourceforge.net/ontologies/MGEDOntology.owl>

¹³ <http://swan.mindinformatics.org/ontology.html>

References

1. Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F., Stein, L.A.: OWL Web Ontology Language Reference. W3C. (February 2004) W3C Recommendation.
2. De Roure, D., Goble, C., Bhagat, J., Cruickshank, D., Goderis, A., Michaelides, D., Newman, D.: myExperiment: Defining the social virtual research environment. In: 4th IEEE International Conference on e-Science, IEEE Press (December 2008) 182–189
3. De Roure, D., Goble, C., Stevens, R.: The design and realisation of the myExperiment virtual research environment for social sharing of workflows. *Future Generation Computer Systems* **25** (February 2009) 561–567
4. Thomas, D., Hansson, D.H., Breedts, L., Clark, M., Davidson, J.D., Gehrtland, J., Schwarz, A.: Agile Web Development with Rails. 2nd edn. Pragmatic Bookshelf (2007)
5. Sahoo, S.S., Halb, W., Hellmann, S., Idehen, K., Thibodeau Jr, T., Auer, S., Ezzat, A.: A survey of current approaches for mapping of relational databases to RDF. Technical report, W3C RDB2RDF Incubator Group (January 2008)
6. Cerbah, F.: Learning highly structured semantic repositories from relational databases - the RDBtoOnto tool. In: Proceedings of the 5th European Semantic Web Conference (ESWC 2008). (June 2008)
7. Cullot, N., Ghawi, R., Yetongnon, K.: DB2OWL: A tool for automatic database-to-ontology mapping. In: Proceedings of the 15th Italian Symposium on Advanced Database Systems (SEBD 2007). (June 2008) 491–494
8. Glaser, H., Jaffri, A., Millard, I.: Managing co-reference on the semantic web. In: WWW2009 Workshop: Linked Data on the Web (LDOW2009). (April 2009)
9. Mons, B.: Which gene did you mean? *BMC Bioinformatics* **6**(142) (June 2005)
10. De Roure, D., Goble, C., Alekseyevs, S., Bechhofer, S., Bhagat, J., Cruickshank, D., Fisher, P., Hull, D., Michaelides, D., Newman, D., Procter, R., Lin, Y., Poschen, M.: Towards open science: The myExperiment approach. *Concurrency and Computation: Practice and Experience* **preprint** (April 2009)
11. Johnston, P., Nelson, M., Sanderson, R., Warner, S.: ORE User Guide - Primer. Open Archives Initiative. (October 2008)
12. Ciccurese, P., Wu, E., Wong, G., Ocana, M., Kinoshita, J., Ruttenberg, A., Clark, T.: The SWAN biomedical discourse ontology. *J. of Biomedical Informatics* **41**(5) (2008) 739–751