A Linked Data Approach to Sharing Workflows and Workflow Results

Marco Roos[1,2], Sean Bechhofer[3], Jun Zhao[4], Paolo Missier[3], David R. Newman[5], David De Roure[6], M. Scott Marshall[2,7]

[1]BioSemantics Group, Department of Human and Clinical Genetics, Leiden University Medical Centre, P.O. Box 9600, 2300 RC Leiden, The Netherlands
[2]Informatics Institute, Faculty of Science, University of Amsterdam, P.O. Box 94323, 1090 GH Amsterdam, The Netherlands
m.roos@lumc.nl
[3]School of Computer Science, The University of Manchester, Manchester, UK
{paolo.missier,sean.bechhofer}@manchester.ac.uk
[4]Jun Zhao, Department of Zoology, University of Oxford.
South Parks Road, Oxford, OX1 3PS
jun.zhao@zoo.ox.ac.uk
[5]School of Electronics and Computer Science, University of Southampton, Highfield Campus, University Road, Southampton SO17 1BJ, UK
drn05r@ecs.soton.ac.uk
[6]Oxford e-Research Centre, University of Oxford, 7 Keble Road, Oxford OX1 3QG, UK
david.deroure@oerc.ox.ac.uk
[7]Department of Medical Statistics and Bioinformatics, Leiden University Medical Centre, P.O. Box 9600, 2300 RC Leiden, The Netherlands
mscottmarshall@gmail.com

Abstract. A bioinformatics analysis pipeline is often highly elaborate, due to the inherent complexity of biological systems and the variety and size of datasets. A digital equivalent of the 'Materials and Methods' section in wet laboratory publications would be highly beneficial to bioinformatics, for evaluating evidence and examining data across related experiments, while introducing the potential to find associated resources and integrate them as data and services. We present initial steps towards preserving bioinformatics 'materials and methods' by exploiting the workflow paradigm for capturing the design of a data analysis pipeline, and RDF to link the workflow, its component services, run-time provenance, and a personalized biological interpretation of the results. An example shows the reproduction of the unique graph of an analysis procedure, its results, provenance, and personal interpretation of a text mining experiment. It links data from Taverna, myExperiment.org, BioCatalogue.org, and ConceptWiki.org. The approach is relatively 'light-weight' and unobtrusive to bioinformatics users.

Keywords: Linked Data, Semantic Web, Digital preservation, Workflow, Provenance, Concept Web

1 Introduction

A commonly used approach to the study of biological systems in the omics era is to integrate information from multiple resources, often in the context of interpreting our own data from an in-house omics experiment (e.g. genome-wide gene expression). The bioinformatics analysis pipeline is therefore usually complex, while the amount of relevant knowledge that could theoretically be considered for a new hypothesis is daunting. With over 19 million biomedical publications in PubMed alone and over a thousand public databases, information overload is a general problem in biology. Although these numbers are impressive, the abundance of information only translates into knowledge gain if we can locate and leverage the knowledge contained in the many distributed resources, including derived data and knowledge extracted by workflows or other computational means. Many have responded to the challenge by aggregating valuable or otherwise thematic data in data warehouses and making the integrated data available on the Web in the form of a knowledge base. However, in all cases, the challenge remains to create a system for the description and subsequent computational discovery of distributed knowledge resources so that they can be incorporated into additional experiments and hypothesis testing.

Not surprisingly, sharing a bioinformatics experiment and its results can be challenging, whether for reuse of its results and its methodology or for peer evaluation. In a networked environment, sharing involves a search process in order to select from a potentially vast number of varied offerings. For wet laboratory experiments, this is supported in particular by the 'Materials and Methods' section of a publication, which describes how the results were obtained. It describes the protocol that was followed, often referring to protocols in earlier publications or in journals and books dedicated to protocols. It describes the specimens and equipment used in enough detail to reproduce the experiment. In many cases, strict nomenclature is imposed by publishers to name, for example, genes and organisms. This makes it easier for peers to understand the experiment, thereby facilitating its review and reuse of its protocol. The Materials and Methods section is considered one of the pillars of experimental biology and probably the most critically assessed part of scientific discourse. A digital version of the Materials and Methods for a bioinformatics experiment would increase its reusability, as well as the rigor by which it can be evaluated. However, a digital equivalent of the 'Materials and Methods' section does not yet exist in bioinformatics. In this paper, we show how a workflow system and web-based information repositories can be used to create the digital equivalent of the Materials and Methods section when we adopt a Semantic Linked Data approach. In the remainder of this paper, we describe the user requirements and their technical counterparts, before describing the components that provide the basis of our approach, supported by a proof of principle in section 3.

1.1 Motivating scenario

To motivate our approach with a scenario we introduce Alice. Alice is interested in performing a bioinformatics experiment to discover proteins that interact with transmembrane proteins, particularly those that can be related somehow to neurodegenerative diseases in which protein aggregates (amyloids) play a significant role (e.g. Huntington's Disease and Alzheimer's Disease). Alice would like to reinvent as little as possible, thus reuse any previously developed analysis pipeline that she can trust to be of the appropriate relevance and quality. Consequently, the typical experiment cycle may contain these four steps:

(i) *Retrieve:* Alice needs to find a previously published analysis pipeline that will suit her needs, and retrieve all the relevant resources (data and methods) for her own analysis.

(ii) *Review:* she will want to review the analysis pipeline before she uses it and study the evidence that led to the interpretation of the data that it previously produced. In theory, the aggregation of metadata associated with the previous experiment should suffice to completely understand the process from input to output to biological interpretation.

(iii) *Repeat, Reuse, Repurpose:* first, Alice would like to repeat a previous analysis to evaluate the process step by step as part of reviewing and validating it. Secondly, Alice would like to be able to run (parts of) an analysis pipeline again for her own purposes, much like bench biologists design new experiments from previously published protocols.

(iv) *Conserve:* when Alice has performed her own analysis, she would like to conserve her design and the association of the analysis with her results, her interpretation, and her initial hypothesis. A bench biologist would keep this type of information in a laboratory journal as the basis for a publication. Alice would like to keep notes on (intermediate) results, the steps that she performed at a particular time, the protocols she used, and any additional information that she may need to support her interpretation of the data. Obviously, the quality of this step determines how effectively Alice's colleague, Bob, would be able to evaluate and reuse Alice's work.

In the next session, we discuss bottlenecks and requirements for performing these steps effectively for bioinformatics experiments.

1.2 Bottlenecks for Evaluating a Bioinformatics Experiment

We identify the following bottlenecks for biologists who wish to be able to evaluate and reuse a bioinformatics experiment:

1. **Retrieve.** Currently, search engines such as Google and NCBI's PubMed are the most common tools to find related work, including methods and (references to) data. This may serve some purposes well enough, but is limited by how precisely we can formulate a search query. In the scenario above Alice will find that it is difficult to find a protein interaction discovery method in the literature using these

tools. Most titles refer to a biological finding rather than the method that was used. She will find that data can often be retrieved on request from the authors or via a public database, but the original analysis pipeline is often not readily available, nor its component parts. Alice will often find it frustrating that her desired method cannot be used independently from the monolithic application in which it is embedded. In this paper, we refer to workflows in myExperiment and Web Services in BioCatalogue to address these issues. Other partial solutions have been developed for bioinformatics, such as BioConductor, popular for developing and sharing statistical analysis methods in the R language [1], or BioMoby, a project that pioneered the use of semantically annotated web services [2].

2. **Review.** While Alice may read the authors' description of a bioinformatics experiment in a publication, she will often find it hard to evaluate its steps. She will not be able to obtain an evidence graph from input to output to biological interpretation, i.e. no data provenance that links between the analysis pipeline and its results is available. The feature-rich (web) application mentioned above is not sufficient to evaluate the underlying computational pipeline. Moreover, additional information that Alice would like to use for her evaluation can be hard to access. For instance, she may want to find which parts in a pipeline were based on other pipelines, which scientists corroborate previous results, or which diseases are associated with the proteins in the result set. There is currently no standardized interface in bioinformatics that makes it possible to query across data, methods and interpretations. In this paper we demonstrate the use of Linked Data and RDF (See Section 1.3), but these are not yet commonly applied in this context.
3. **Repeat, Reuse, Repurpose.** It is often difficult to repeat a Bioinformatics experiment. As mentioned, component parts may not be available for a new application and even when a client application is available to rerun the full pipeline, the underlying databases may have been updated or computational methods improved. This cannot be completely controlled when applications are built on 3rd party resources, but Alice would be helped greatly if she was notified of changes such that she can take these into account when rerunning a method. Workflows built from Web Services seem address part of these bottlenecks.
4. **Conservation.** In the laboratory, the most generally accepted method for conservation of methods, data, and interpretation is still traditionally through publications and lab journals. Many publishers offer the option of supplying additional digital information, but the quality of this 'supplementary information' varies and it is not usually computationally accessible due to a lack of standardization, nor does it provide a way to link the analysis pipeline, its results, and associated metadata. For example, although it is common practice to upload raw data from microarray studies to ArrayExpress and GEO, the lists of differentially expressed genes commonly referred to by articles are not disclosed with the raw data. However, it is precisely these lists that are the subject of discussion in any associated articles. Alice will find that she has no direct way to associate her notes (her annotations) with her analysis design and its results.

New Web2.0-inspired applications provide alternative ways to digitally conserve analysis designs (myExperiment; [3]), their component parts (BioCatalogue; [4]), and the concepts used in biological hypotheses and personal notes (ConceptWiki[1]; [5][2]). In this paper, we describe how we use RDF to link some of these resources together to create a comprehensive digital resource that describes the 'materials and methods' of a bioinformatics experiment, and we discuss how this addresses Alice's bottlenecks. First, we identify these additional user requirements:

1. **Comprehensive.** Alice would be helped in reviewing a previous analysis if she would be able to query a comprehensive 'warehouse' of information about the methods and the data associated with an experiment. For instance, she may want to look at alternative gene names, related diseases, or author names and affiliations.
2. **Light Weight.** While Alice would like to query a comprehensive 'warehouse' to perform an extensive review, she would not want to spend substantial effort to build this warehouse herself. Moreover, she would not want to do so for her own analysis that Bob will review.
3. **Transparent.** The technology for digital conservation relies on semantic annotation of the components of an experiment and its results. However, this should not interfere with the design of the bioinformatics analysis. In fact, Alice should feel supported by it, for instance by implementing it as a tool that helps her keep a laboratory journal. The activities that result in a digital version of materials and methods should ideally be part of her routine research activity.
4. **Personal.** In general, reusing a community consensus model to annotate the results of an analysis will help Alice and Bob to interpret her results. However, Alice's work is cutting-edge, so she has a personal view of her bioinformatics experiment that is reflected in her hypothesis and data interpretation. Therefore, Alice requires the ability to use the most appropriate model for her annotations, and the ability to extend an existing model with concepts that she is missing.
5. **Shared terminology, Identity and Reference.** In biological discourse, various 'nomenclatures' (e.g. for species or gene names) are used to resolve ambiguity. Also for a bioinformatics analysis we depend on unambiguous and unique identifiers for the objects in our digital materials and methods. In this paper, we use the Concept Web, a new part of the Semantic Web that aims to be a world-wide resource of disambiguated (biological) concepts, machine readable through RDF and identified by universally unique identifiers.

1.3 Semantic Web, RDF and Linked Data

The Semantic Web as described by W3C[3] is about providing common formats for integration and combination of data drawn from diverse sources. The Semantic Web aims to lift us from a web of pages or resources with data intended solely for human

[1] http://ConceptWiki.org
[2] Originally based on WikiProfessional technology: http://wikiprofessional.org
[3] http://www.w3.org/2001/sw/

consumption to a "web of data", with this data explicitly exposed, rather than locked away inside particular applications.

The Resource Description Framework (RDF[4]) is seen as a key technology in the publication of the web of data, including data from the Life Sciences [6, 7]. RDF provides a common triple-based data model for publication of data. It is indeed increasingly used to expose data sets and resources as RDF graphs. SPARQL[5] provides a language for querying graph patterns within RDF graphs, and also defines a protocol that describes how queries can be conveyed to a SPARQL "endpoint", a service that processes SPARQL queries. SPARQL thus enables the query of RDF data sets and provides a common infrastructure on which to build applications.

An approach that is steadily growing in popularity is that of Linked Data [8, 9]. Linked Data is a set of guidelines or best practices that have been introduced in order to facilitate the exposure and connection of different data sets. The Linked Data approach relies heavily on RDF and the use of URIs to identify objects or concepts that are being described. Linked Data advocates the following principles:

1. Use URIs to identify objects/concepts, in particular use HTTP URIs which are then dereferencable.
2. Provide useful information when those URIs are dereferenced, ideally using standard formats and representations (e.g. RDF)
3. Provide links to other URIs, so that applications can discover more.

The adoption of these guidelines for the publication of data enables the integration of data sets from a wide range of domains, with significant efforts in the life sciences. Key issues facing the Linked Data approach include the provision of common, shared identifiers for the objects that are being described -- the use of common URIs drives the "linking" in Linked Data. Ensuring that applications and datasets use common identifiers is thus crucial in facilitating this linking. Initiatives such as Shared Names[6], Okkam[7] and the Concept Web[8] (as discussed later) are aiming to provide URIs for publicly available records. Authoritative resources such as UniProt, PubMed and EntrezGene are also being exposed as RDF via SPARQL endpoints by projects including Linked Life Data[9] or Bio2RDF[10].

In the context of our scenario, there are additional objects that can be identified and linked together, as discussed below. These include the *workflows* that are used to process the data, the *services* that are used within those workflows, the *researchers* who conduct the research and the *outputs* (papers, presentations etc) that those researchers produce. Exposing all of these resources as Linked Data will provide a

[4] http://www.w3.org/RDF/
[5] http://www.w3.org/TR/rdf-sparql-query/
[6] http://sharedname.org/
[7] http://www.okkam.org/
[8] http://www.conceptweballiance.org/
[9] http://linkedlifedata.com/
[10] http://bio2rdf.org/

rich, connected space facilitating discovery, analysis and reuse of digital materials and methods.

2 Resources for Digital Materials and Methods

Here we describe how the Linked Data principles are used to aggregate the resources that Alice could use for (i) retrieving a previously constructed pipeline for protein discovery, its component parts, and associated documentation (myExperiment, BioCatalogue), (ii) reviewing the analysis and its results (Taverna workflow provenance with domain specific extensions), (iii) repeating the analysis for her own purposes, and (iv) classifying the results: protein interactions found by a text mining workflow (Taverna+AIDA plugin). Only a limited number of additional links are necessary to create a new aggregation that represents the digital materials and methods of Alice's experiment. We demonstrate this with an example in section 3.

2.1 RDF: The Model for Linked Data and Comprehensive, yet Light-weight Coverage of Experiment-related Data

Our framework of choice for digitally conserving a computational analysis encompassing hypothesis, provenance, workflow(s), services, data, and interpretation is based on RDF (section 1.3). Many applications have started exposing their data on the web via RDF, making their resources part of the Linked Open Data cloud. This can be done either by a SPARQL endpoint or by providing RDF as a machine readable alternative to the data presented on a web page. This includes the resources that we have identified as useful sources for our digital Materials and Methods: Taverna, myExperiment, BioCatalogue, and the Concept Web. With a minimal number of links between these sources, Alice is provided with a comprehensive amount of metadata about an experiment.

2.2 myExperiment and BioCatalogue: Repositories for Digital Protocols and their Components

While the workflow paradigm provides a useful way to formalise an analysis pipeline, myExperiment.org provides a repository to share and publish these artefacts on the Web [10]. Additional documentation (tags, comments) can be provided by the owner of a workflow or users of myExperiment. This facilitates their discovery and reuse. In turn, BioCatalogue provides a registry for the components of a workflow, in particular Web Services [4]. Similar to myExperiment, BioCatalogue enables registered users to provide documentation and tag contents, again facilitating their discovery. Both resources provide a REST API and URLs for every object that they contain. Conequently, myExperiment and BioCatalogue are sources of identifiers for use in bioinformatics publications. Versioning and attribution features ensure that specific adaptations of a workflow can be referenced. Attribution allows Bob to link to Alice's

workflow and acknowledge her. When Alice also attributes a workflow, then these links implicitly create a chain of references to the origins of a workflow. Finally, we mention myExperiment 'packs': aggregations of (references for) resources both inside and outside of myExperiment. This makes myExperiment a provider of persistent and structured supplemental information. How can we use myExperiment and BioCatalogue to link to Alice's experimental results and create the digital version of her Materials and Methods? MyExperiment also exposes its content as RDF [11]. The motivation is to make the content of myExperiment part of Linked Data, allow it to be linked to other resources and be queried via a SPARQL endpoint. This will allow Alice to retrieve information from the Web of Data starting from a myExperiment pack. The semantic model that was used for myExperiment supports its core features. It represents the social model behind myExperiment and the model that facilitates the management and sharing of workflows and associated components for other users. This 'e-Research' model is extensible such that it can be linked to additional domain specific models. The most straightforward part of the myExperiment semantic model is the representation of the myExperiment mySQL schema in OWL DL. The Simple Network Access Rights Management (SNARM[11]) ontology was used to capture the sharing model of myExperiment. For representing the social content of myExperiment several ontologies were reused: Dublin Core[12], Friend of a Friend[13], Semantically Inter-linked Online Communities (SIOC[14]), and the Open Archives Initiative's Object Reuse and Exchange ontologies/schemata (OAI/ORE[15]). These shared ontologies facilitate co-reference resolution, which is one of the major tasks on the Semantic Web. It makes it easier to understand the purpose of the classes and relations and facilitates access to semantic data outside of myExperiment. The users of the user interface are never confronted with the full extent of these ontologies. Exposing the content of myExperiment as Linked Data on the web allows Alice to define SPARQL queries for typical Materials and Methods questions such as 'Who did what and when?', or 'Whose work was this workflow based on?'. Moreover, the relatively straightforward action for Alice to upload and publish her workflow on myExperiment provides Bob, a potential new user, additional metadata to investigate. At the time of writing BioCatalogue does not yet expose its content as RDF. For our example in section 3 we used myExperiment RDF as a template to create a mock version of BioCatalogue RDF.

2.3 Workflow and Provenance

Workflows are the most common type of object that Alice finds on myExperiment for reusing in her own work. Workflows are formal and executable models of computational protocols for data analysis experiments. Alice can review the design of

[11] http://rdf.myexperiment.org/ontologies/snarm/
[12] http://dublincore.org/
[13] http://www.foaf-project.org/
[14] http://sioc-project.org/
[15] http://www.openarchives.org/ore/

a workflow, similar to how she would evaluate a protocol from a laboratory manual. However, the best way to review an experiment before using it for one's own purposes is to evaluate the results that it produced step by step and the personal annotations that the first user of the workflow provided while he/she was running it. In comparison, if Alice was to reuse a wet laboratory protocol by bench biologist Chris, then his laboratory notes made while he was performing the protocol would be the most valuable. First, they contain what was actually done in relation to the results at a particular point in time. Secondly, it contains Chris' personal annotations on how the results should be interpreted. Therefore, capturing a detailed trace, the *provenance*, of each workflow execution (a "run") linked with personal annotations represents a step forward in the direction of recording materials and methods in machine processable form. The Taverna workflow system persistently stores the provenance of workflow runs (for example, the execution of Alice's experiment) and makes it available to scientists for evaluation. At any later time Bob can query and analyse Alice's results. Taverna adopts a semantic data model to represent provenance. The model is specified as an OWL ontology, called *Janus*. Provenance traces are RDF graphs [12]. The concepts in Janus describe workflow tasks as well as the data that they consume and produce, while the provenance graph captures the actual tasks and the data transformations that they produced during a workflow run.

The choice of a semantic model is designed to facilitate the semantic annotation of provenance graphs with domain-specific concepts, such as those found on the Concept Wiki. When provenance is first recorded, the provenance graphs are "domain-agnostic" and semantics-free, but their grounding in RDF and OWL makes it easy to add annotations whenever they become available, and to integrate with the broader Web of Linked Open Data [13]. Such integration involves mapping data elements in the provenance graph to data that is published elsewhere in the Web of Data, making it possible for queries to seamlessly include conditions on properties of the data that were not explicitly represented in the original graph. Henceforth, without bloating the original provenance produced by the workflow enactor, a comprehensive graph can be obtained via meaningful relations on the Semantic Web.

Janus achieves the required linking by reusing a number of shared ontologies found on the web. Formally, Janus is an extension of Provenir [14], which itself extends concepts from the Basic Formal Ontology (BFO[16]). Provenir is an upper-level reference model for capturing provenance, including concepts such as data, process and agent, and several relations such as for partonomy, precedence, and causality. Janus extends Provenir to include terms from the Life Sciences domain. For example, four ontologies were chosen for case studies in genomics from the almost 200 publicly shared models that are available via the National Centre for Biomedical Ontologies (NCBO; [15]): BioPAX[17], the National Cancer Institute Thesaurus[18], the Foundational Model of Anatomy[19], and the Sequence Ontology[20]. Doing so following

[16] http://www.ifomis.org/bfo
[17] http://www.biopax.org/
[18] http://ncit.nci.nih.gov/
[19] http://sig.biostr.washington.edu/projects/fm/
[20] http://www.sequenceontology.org/

Linked Data conventions allow Alice and Bob to ask useful biological questions about interacting biological molecules from KEGG, Reactome, and BioCyc databases [16]. As such, provenance becomes the core of a comprehensive digital resource of materials and methods for biologists to evaluate and reuse.

2.4 Concept Web: Repository for Uniquely Identified Concepts, their Relations and their Evidence

A new approach to providing common identifiers for important terms in scientific discourse is proposed by the Concept Web Alliance [17]. Inspired by the success of Wikipedia, it 'calls upon a million minds' to create and curate a universal resource of disambiguated concepts and basic relations between them [5]. In line with a wiki approach, scientists can register new concepts and improve the information associated with them. Initial content is supplied by terminology resources such as UMLS, UniProt, and the ontologies that can be obtained from NCBO's bioportal [15]. Relations can be aggregated to form so-called 'nano-publications' [17]. 'Malaria' and 'mosquito' are example concepts, while 'Malaria *is caused by* mosquitos *as discovered by* Charles Laveran *in* 1880' could be a nano-publication including a trace to evidence. Each concept, relation, and nano-publication will have its own universally unique identifier that is persistent and immutable over time. Therefore, Alice and her peers can use Concept Web identifiers as stable references to their data instead of, for instance, gene names, which can change. Because the Concept Web is also part of the Semantic Web and exposes its content as RDF, it is a unique source of identifiers for use on the Web of Data. In our proof of principle, we will use concepts from the Concept Web as our point of reference for all digital objects except those from myExperiment and BioCatalogue. These resources already claim that their URLs are persistent and universally unique.

3 Proof of principle

3.1 Linking Experimental Results and Evidence (Taverna Provenance), Personal Interpretation (AIDA plugin), Digital Protocol (myExperiment) and its Components (BioCatalogue), in terms of Biological Concepts (ConceptWiki)

Here we show how we obtain a snapshot out of the digital, machine readable Materials and Methods as a result of Alice running her workflow. Bob would like to use this information to review how Alice obtained these results, and study some additional information about these results. Our example is derived from the workflow that Alice was using for protein discovery. We have used a number of resources that Bob would need to satisfy his information needs. When resources follow the Linked Data principles we require only a minimal number of relations to embed Alice's

workflow results in a large network of references. Therefore, this solution is light-weight, but still comprehensive. The following Linked Data resources were used:

1. Taverna provenance: exposed as RDF using *Janus* (section 2.1)
2. myExperiment: a provisional RDF document for the protein discovery workflow was obtained from the myExperiment development server (section 2.2)
3. BioCatalogue: we created a mock RDF document using myExperiment RDF data as example. A RDF interface similar to that of myExperiment is planned (Jiten Bhagat, personal communication)
4. ConceptWiki: provisional RDF documents were obtained from the ConceptWiki development server. We created new concepts via the ConceptWiki interface to obtain universally unique identifiers for the creator of the workflow and services. Ideally, myExperiment and BioCatalogue would use these as identifiers as well.
5. UniProt: the RDF document for our example protein was obtained from the RDF interface of the main UniProt web site.

To link these resources, we used properties from the following ontologies:

1. A Workflow ontology previously created for structuring data from a workflow [18]
2. A mapping ontology for mapping between a (text mining) process and biological results [18]
3. The Semantic Web Applications in Neuromedicine (SWAN; [19]) ontology version 1.2[21]
4. The Relation Ontology (RO) from the OBO Foundry [20]
5. The Dublin Core (DC) meta-thesaurus.

The following links were made (see appendix for the commented RDF):

Between Taverna provenance and reference resources:

- a workflow run *is a run of* a workflow on myExperiment
- an executed processor *is a run of* a service on BioCatalogue
- a workflow result *is the result of* a service run
- a workflow result *refers to* a concept on the Concept Web

Between BioCatalogue, myExperiment, Concept Web, and UniProt

- a service in BioCatalogue *is an element of* a workflow on myExperiment
- a workflow *is created by* a user who is identified on the Concept Web
- a service in BioCatalogue *has a creator* who is identified on the Concept Web
- a Concept Web entry *cites* a UniProt entry and vice versa

Missing links

- a processor_exec *participates in* a workflow run

[21] http://swan.mindinformatics.org/ontology.html

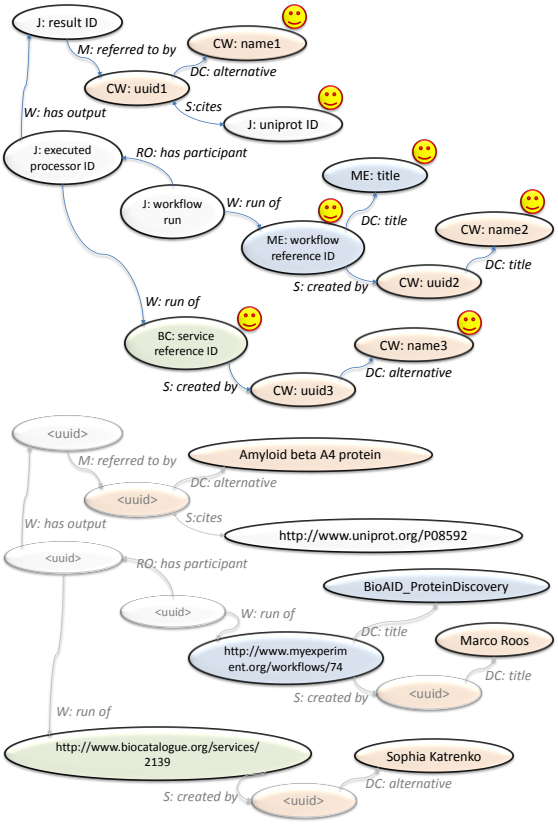

Fig. 1. Top: graphic representation of an evidence query. The smiley symbols indicate which elements could be in the output for human consumption. Bottom: graphic representation of the result of the evidence query. '<uuid>' indicates a universally unique identifier that is provided by any of the resources. J: Janus Provenance Ontology; M: mapping ontology to relate the (text mining) process to biological concepts; CW: Concept Web (the Concept Wiki is the human GUI); S: SWAN 'Semantic Web Applications in Neuromedicine' ontology; W: Workflow ontology; RO: OBO Relations ontology; ME: myExperiment; DC: Dublin Core; BC: BioCatalogue. See the appendix for the SPARQL representation and its results.

When the results of the Protein Discovery workflow are linked to the resources that were used to create them, and to the resources that refer to and describe her results, Bob can obtain a comprehensive view to review for instance how she obtained her results, who were the people responsible for the methods and the workflow, and the 'linked in' identifiers that Bob could use to further interpret Alice's results. Bob can obtain this information by querying the RDF graph. We demonstrate this by a 'materials and methods' query (figure 1). First we created a SPARQL endpoint by uploading the RDF documents described above to our Sesame RDF repository. For our proof of concept we focussed on one of the workflow results, the protein 'Amyloid beta A4 protein' (UniProt identifier: P08592[22]). The graph pattern in figure 1 shows how evidence for a workflow result was queried (see the appendix for the full SPARQL query). Bob would be able to study this evidence, and continue querying for new information. For instance, Bob could write a query that retrieves all literature citations contained in a UniProt RDF document that was ultimately linked to Alice's workflow result as created by Taverna's provenance engine.

[22] http://www.uniprot.org/uniprot/P08592

4 Discussion and Conclusion

How do the components that we have presented alleviate the bottlenecks that Alice and Bob face in their research? What are the potentially new bottlenecks that we have not solved?

Retrieve. Alice and Bob are supported in their retrieval task in two ways. First, myExperiment.org, BioCatalogue, and the ConceptWiki index their content such that it can be searched through keyword searches. 'Materials and Methods' aggregates stored as myExperiment packs (e.g. packs 82[23] and 58[24]) can thus be found via a familiar search interface. Secondly, when data is exposed as RDF, the Semantic Web query language SPARQL can be used to retrieve precise graphs from the Web of Data. Semantic search can be further facilitated for Alice and Bob if we can hide the SPARQL syntax and the complexity of federated queries through a familiar keyword-based search interface that incorporates auto-completion and browsing of related concepts. As a query language, SPARQL is meant for use by developers so a sufficiently user friendly interface would supply common search patterns, built on a set of SPARQL queries.

Review. The review process is supported, because by using RDF to expose data and to link data it is now possible for Bob to query the complete evidence graph from hypothesis to input to experiment to output to interpretation. In section 3, we demonstrated this principle by retrieving the service that produced one of the proteins in our result set and the creator of that service. Bob can further explore the meaning of Alice's results by exploring any additional links contained in our RDF resources. For instance, we can retrieve extra information from UniProt.

Repeat, reuse, repurpose. Workflows are particularly useful to repeat, reuse, or repurpose a bioinformatics analysis pipeline. A workflow created with a workflow system like Taverna can be reused in new designs. Semantic annotation as facilitated by the AIDA plugin, which gives extra information about the intent of the workflow, which in turn makes it easier to reuse. The particularly tricky bottleneck by changes in services or their underlying data can be partially addressed by a forthcoming feature of BioCatalogue that will indicate when a service has changed its interface or its behavior. It also indicates whether the service is up and running.

Conserve. Embedding data and models in semantic models exposed as RDF/Linked Data provides Alice and Bob with an alternative way to publish and share information. Using identifiers from the Concept Web further lowers the threshold to link information across the web, and to study those links. When myExperiment packs can be accessed via RDF as Research Objects with a consistent interface across the world we have successfully created to new paradigm for scientific publications.

[23] http://www.myexperiment.org/packs/82
[24] http://www.myexperiment.org/packs/58

4.1 Research Objects for Publication

The myExperiment 'pack' provides a mechanism for bundling together a collection of resources. The pack is relatively simple in terms of structure, however, essentially providing a "zip file" containing resources, or references to resources. The pack itself can then be annotated with appropriate metadata and shared through myExperiment. As discussed above, the relationships between the resources involved in an experiment (data, methods, results, provenance) is much richer than a simple collection. Packs can thus be seen as a first approximation to a 'Research Object' (RO), a mechanism for publishing reproducible research that is shared on the Web [21]. An RO provides a container for the aggregation of resources, along with information about the relationships between those resources. For Alice, an RO contains all the artefacts that Alice would consider a complete 'experiment', while for her peer Bob it contains everything he needs to reproduce the experiment. As discussed in [21], ROs then provide support for reusability, allowing replay of experiments, repetition of experiments and repurposing of experiments, building on the methods and materials employed. As future work, ROs that have been described with ontological annotations could strengthen the validation and review part of our scenario and provide a self-contained set of procedures and accompanying resources.

5 Acknowledgements

We thank Katy Wolstencroft and Andrew Gibson for suggestions and critically reading the manuscript, and the teams of the myGrid project, myExperiment, BioCatalogue, the Concept Web Alliance and the Netherlands BioInformatics Centre (NBIC) for their support.

6 References

1. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y.H., Zhang, J.: Bioconductor: open software development for computational biology and bioinformatics. Genome Biology. 5, R80 (2004).
2. Wilkinson, M.D., Links, M.: BioMOBY: an open source biological web services proposal. Briefings in Bioinformatics. 3, 331-341 (2002).
3. Goble, C.A., Bhagat, J., Aleksejevs, S., Cruickshank, D., Michaelides, D., Newman, D., Borkum, M., Bechhofer, S., Roos, M., Li, P., De Roure, D.: myExperiment: a repository and social network for the sharing of bioinformatics workflows. Nucleic Acids Research. (2010).
4. Bhagat, J., Tanoh, F., Nzuobontane, E., Laurent, T., Orlowski, J., Roos, M., Wolstencroft, K., Aleksejevs, S., Stevens, R., Pettifer, S., Lopez, R., Goble, C.A.: BioCatalogue: a universal catalogue of web services for the life sciences. Nucleic Acids Research. (2010).
5. Mons, B., Ashburner, M., Chichester, C., van Mulligen, E., Weeber, M., den Dunnen, J., van Ommen, G.J., Musen, M., Cockerill, M., Hermjakob, H., Mons, A., Packer, A., Pacheco, R.,

Lewis, S., Berkeley, A., Melton, W., Barris, N., Wales, J., Meijssen, G., Moeller, E., Roes, P.J., Borner, K., Bairoch, A.: Calling on a million minds for community annotation in WikiProteins. Genome biology. 9, R89 (2008).
6. Neumann, E., Miller, E., Wilbanks, J.: What the semantic web could do for the life sciences. Drug Discovery Today: BIOSILICO. 2, 228-236 (2004).
7. Marshall, M., Post, L., Roos, M., Breit, T.: Using Semantic Web Tools to Integrate Experimental Measurement Data on Our Own Terms. On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops. pp. 688, 679 %U http://dx.doi.org/10.1007/11915034_92 (2006).
8. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data. International Journal on Semantic Web and Information Systems. 5, (2009).
9. Bizer, C., Heath, T., Idehen, K., Berners-Lee, T.: Linked data on the web (LDOW2008). Proceeding of the 17th international conference on World Wide Web - WWW '08. p. 1265, Beijing, China (2008).
10. De Roure, D., Goble, C., Bhagat, J., Cruickshank, D., Goderis, A., Michaelides, D., Newman, D.: myExperiment: Defining the Social Virtual Research Environment. (2008).
11. Newman, D., Bechhofer, S., Roure, D.C.D.: myExperiment: An Ontology for e-Research. Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009). , Washington DC, USA (2009).
12. Missier, P., Sahoo, S., Zhao, J., Goble, C.A., Sheth, A.: Janus : from workflows to semantic provenance and linked open data. Proceedings of The third International Provenance and Annotation Workshop. , Troy, NY, U.S.A. (2010).
13. Zhao, J., Miles, A., Klyne, G., Shotton, D.: Linked data and provenance in biological data webs. Briefings in Bioinformatics. 10, 139-152 (2009).
14. Sahoo, S., Sheth, A.: Provenir ontology: Towards a Framework for eScience Provenance Management. Microsoft eScience Workshop. , Pittsburgh, PA, USA (2009).
15. Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M., Chute, C.G., Musen, M.A.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research. 37, W170-173 (2009).
16. Luciano, J.S., Stevens, R.D.: e-Science and biological pathway semantics. BMC Bioinformatics. 8 Suppl 3, S3 (2007).
17. Mons, B., Velterop, J.: Nano-Publication in the e-science era. Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009). p. 14CEUR-WS, Washington DC, USA.
18. Roos, M., Marshall, M.S., Gibson, A.P., Schuemie, M., Meij, E., Katrenko, S., van Hage, W.R., Krommydas, K., Adriaans, P.W.: Structuring and extracting knowledge for the support of hypothesis generation in molecular biology. BMC bioinformatics. 10 Suppl 10, S9 (2009).
19. Clark, T., Kinoshita, J.: Alzforum and SWAN: the present and future of scientific web communities. Briefings in bioinformatics. 8, 163-71 (2007).
20. Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L., Rosse, C.: Relations in biomedical ontologies. Genome Biology. 6, R46 (2005).
21. S. Bechhofer, D. De Roure, M. Gamble, C. Goble, and I. Buchan. Research Objects: Towards Exchange and Reuse of Digital Knowledge. The Future of the Web for Collaborative Science (FWCS 2010), Workshop at WWW2010, Raleigh NC, 2010 Online Version: http://precedings.nature.com/documents/4626/version/1