Workflow forever: Semantic Web Semantic models and tools for preserving and digitally publishing computational experiments

[Link to publication record in Manchester Research Explorer](#)

Workflow forever: Semantic Web Semantic models and tools for preserving and digitally publishing computational experiments

Kristina Hettne
Department of Human Genetics
Leiden University Medical
Center, Leiden, NL
k.m.hettne@lumc.nl

Khalid Belhajjame
School of Computer Science
University of Manchester
Manchester, UK
khalidb@cs.man.ac.uk

Marco Roos
Department of Human Genetics
Leiden University Medical
Center, Leiden, NL
m.roos@lumc.nl

Stian Soiland-Reyes
School of Computer Science
University of Manchester
Manchester, UK
soiland-reyes@cs.manchester.ac.uk

Matthew Gamble
School of Computer Science
University of Manchester
Manchester, UK
m.gamble@cs.man.ac.uk

Oscar Corcho
Universidad Politécnica de Madrid,
Madrid, ES
ocorcho@fi.upm.es

Graham Klyne
Department of Zoology
University of Oxford
Oxford, UK
graham.klyne@zoo.ox.ac.uk

Sean Bechhofer
School of Computer Science
University of Manchester
Manchester, UK
sean.bechhofer@cs.man.ac.uk

ABSTRACT

One of the main challenges for biomedical research lies in the integrative study of large and increasingly complex combinations of data in order to understand molecular mechanisms, for instance to explain the onset and progression of human diseases. Computer-assisted methodology is needed to perform these studies, posing new challenges for upholding scientific quality standards for the reproducibility of science. This pertains to the preservation of the 'materials and methods' of computational experiments as a record of the evidence for the biological interpretation of their results. We argue that traditional journal publication is no longer sufficient, and propose a methodology based on the workflow paradigm, Semantic Web models, and Digital Library infrastructure.

Our primary goal is to enable the preservation of the necessary and sufficient information for researchers to understand the steps of a computational experiment that led to new biological insight, at any point in the future. Central to our approach is the development of a 'Research Object' (RO) model that captures this information for preservation, publication and acknowledgement. We adopted a combination of a Semantic Web and Digital Library approach for the representation and publication of such a model. The RO model can be viewed as an artifact that aggregates and annotates a number of resources that are used and/or produced in a given scientific investigation. The figure below

(Figure 1) illustrates a high level description of the elements that are needed to specify a research object.
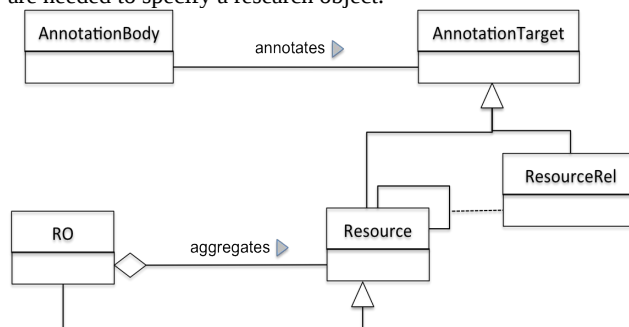

Figure 1: Elements needed to specify a research object

A resource can be a workflow, web service, document, data item, data set, workflow run, software or a research object. Instead of building a new model, we use the Object Reuse and Exchange (OAI-ORE)[1] for specifying aggregation of resources, and Annotation Ontology (AO)[2] for their annotations. ORE[3] defines standards for the description and exchange of aggregations of Web resources. For example, a Research Object can be defined as an ore:Aggregation, and an ore:ResourceMap can be used to describe the research object and its constituent resources (ro:ResearchObject a owl:Class; rdfs:subClassOf ore:Aggregation). Annotations in a RO are specified using the Annotation Ontology[4], which provides a common model for

---

[1] http://www.openarchives.org/ore/

[2] http://code.google.com/p/annotation-ontology

[3] http://www.openarchives.org/ore/1.0/toc.html

[4] http://code.google.com/p/annotation-ontology/wiki/UnderstandingAO

document metadata, typically for annotating electronic documents or parts of electronic documents. Together with domain-specific vocabularies that extend the generic RO model we can specifically annotate the roles of the individual resources. We aim to develop tooling that facilitates annotation at each step of the research cycle, harvesting metadata from users in small steps.

We present an example of an instantiated prototype RO in the context of a study of Metabolic Syndrome, for which we perform computational experiments that help interpret Genome Wide Association Data by using a special text mining method [1]. While we conceived the experiment, designed and performed it, we populated the prototype RO model and annotated the entities to describe their role and their interrelationships. For instance, we defined that 'a particular ranked list of candidate biological processes was produced by a particular workflow run', for which we assert that 'this particular workflow run is a run instance of a specific GWAS Interpretation Workflow', and 'a specific Text Mining Web Service is used in this particular GWAS Interpretation Workflow', while 'a particular GWAS data set is input to the workflow run'. We also defined that the 'RO is created by Kristina Hettne', 'created at a particular time and date', 'motivated by a particular hypothesis', and 'the result is interpretated through a particular change in the hypothesis'. ROs can also refer to previous work of which the output was used in the experiment. ROs may be related to each other and other resources, which can create a graph of scientific progress.

The results presented here are the outcomes of the EU FP7 project 'Wf4Ever that aims to provide tools and recommendations for digitally preserving computational experiments.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Data sharing; H.5.3 [**Group and Organization Interfaces**]: Collaborative computing

2.   REFERENCES

[1] Jelier R, 't Hoen PA, Sterrenburg E, den Dunnen JT, van Ommen GB, Kors JA, Mons B. 2003. Literature-aided meta-analysis of microarray data: a compendium study on muscle development and disease. *BMC Bioinformatics.* 9 (Jun. 2008), 291.