Proceedings of the Workshop on Semantic Publishing (SePublica 2012)

9th Extended Semantic Web Conference Hersonissos, Crete, Greece, May 28, 2012

edited by Frank Van Harmelen, Alexander García Castro, Christoph Lange, and Benjamin Good

May 28, 2012

Preface

This volume contains the papers presented at SePublica 2012 (`http://sepublica.mywikipaper.org`): International Workshop on the Future of Scholarly Communication and Scientific Publishing held on May 28, 2012 in Hersonissos, Crete, Greece.

There were 9 submissions. Each submission was reviewed by at least 2, and on the average 2.7, program committee members. The committee decided to accept 8 papers.

We would like to thank our peer reviewers for carefully reviewing the submissions and giving constructive feedback.

This proceedings volume has been generated with EasyChair, which made this task really convenient.

April 15, 2012
Bremen

Frank Van Harmelen
Alexander García Castro
Christoph Lange
Benjamin Good

Program Committee

Paolo Ciccarese	Harvard Medical School & Massachusetts General Hospital
Sudeshna Das	Harvard University
Anita De Waard	Utrecht University
Michael Dreusicke	PAUX Technologies GmbH
Kai Eckert	Mannheim University Library
Alexander Garcia	Florida State University Guest Professor
Leyla Jael García Castro	Universitaet der Bundeswehr
Benjamin Good	TSRI
Tudor Groza	School of ITEE, The University of Queensland
Krzysztof Janowicz	University of California, Santa Barbara
Michael Kohlhase	KWARC
Sebastian Kruk	Knowledge Hives sp. z o.o.
Christoph Lange	University of Birmingham, University of Bremen, Jacobs University Bremen
Philippe Cudré-Mauroux	EPFL
Robert Morris	DCR Consulting, LLC, UMASS-Boston, and Harvard University
Steve Pettifer	The University of Manchester
Greg Riccardi	Florida State University
Jodi Schneider	DERI, NUI Galway
Frank Van Harmelen	vu.nl
Trish Whetzel	Stanford University
Jun Zhao	University of Oxford

Additional Reviewers

R

Ritze, Dominique

Contents

Workflow-Centric Research Objects: First Class Citizens in Scholarly Discourse

Khalid Belhajjame[1], Oscar Corcho[2], Daniel Garijo[2], Jun Zhao[4], Paolo Missier[9], David Newman[5], Raúl Palma[6], Sean Bechhofer[1], Esteban García Cuesta[3], José Manuel Gómez-Pérez[3], Graham Klyne[4], Kevin Page[4], Marco Roos[7], José Enrique Ruiz[8], Stian Soiland-Reyes[1], Lourdes Verdes-Montenegro[8], David De Roure[4], Carole A. Goble[1]

[1] University of Manchester, UK. [2] Ontology Engineering Group, Universidad Politécnica de Madrid, Spain. [3] iSOCO, Spain. [4] University of Oxford, UK. [6] Poznan Supercomputing and Networking Center, Poland. [7] Leiden University Medical Centre, Netherlands. [8] Instituto de Astrofísica de Andalucía, CSIC, Spain. [9] Newcastle University, UK.
khalidb@cs.man.ac.uk

Abstract. A workflow-centric research object bundles a workflow, the provenance of the results obtained by its enactment, other digital objects that are relevant for the experiment (papers, datasets, etc.), and annotations that semantically describe all these objects. In this paper, we propose a model to specify workflow-centric research objects, and show how the model can be grounded using semantic technologies and existing vocabularies, in particular the Object Reuse and Exchange (ORE) model and the Annotation Ontology (AO). We describe the life-cycle of a research object, which resembles the life-cycle of a scientific experiment.

1 Introduction

Scientific workflows are used to describe series of structured activities and computations that arise in scientific problem-solving, providing scientists from virtually any discipline with a means to specify and enact their experiments [3]. From a computational perspective, such experiments (workflows) can be defined as directed acyclic graphs where the nodes correspond to analysis operations, which can be supplied locally or by third party web services, and where the edges specify the flow of data between those operations.

Besides being useful to describe and execute computations, workflows also allow encoding of scientific methods and know-how. Hence they are valuable objects from a scholarly point of view, for several reasons: (i) to allow assessment of the reproducibility of results; (ii) to be reused by the same or by a different scientist; (iii) to be repurposed for other goals than those for which it was originally built; (iv) to validate the method that led to a new scientific insight; (v) to serve as *live-tutorials*, exposing how to take advantage of existing data infrastructure, etc. This follows a trend that can be observed in disciplines such

as Biology and Astronomy, with other types of objects, such as databases, increasingly becoming part of the research outcomes of an individual or a group, and hence also being shared, cited, reused, versioned, etc. [11]

However, the use of workflow specifications on their own does not guarantee to support reusability, shareability, reproducibility, or better understanding of scientific methods. Workflow environment tools evolve across the years, or they may even disappear. The services and tools used by the workflow may change or evolve too. Finally, the data used by the workflow may be updated or no longer available. To overcome these issues, additional information may be needed. This includes annotations to describe the operations performed by the workflow; annotations to provide details like authors, versions, citations, etc.; links to other resources, such as the provenance of the results obtained by executing the workflow, datasets used as input, etc.. Such additional annotations enable a comprehensive view of the experiment, and encourage inspection of the different elements of that experiment, providing the scientist with a picture of the strengths and weaknesses of the digital experiment in relation to decay, adaptability, stability, etc.

These richly annotation objects are what we call workflow-centric research objects. The notion of Research Object has been introduced in previous work [20, 19, 1] – here we focus on Research Objects that encapsulate scientific workflows (hence workflow-centric). In particular, we build on earlier work on myExperiment *packs*, which are bundles that contain elements such as workflows, documents and presentations [15]. Other related work is presented in Section 2. In this paper we extend that work making the following contributions: we present a model for specifying workflow-centric research objects (Section 3), and show how it is grounded using semantic technologies; and we characterise and define their lifecycle, illustrating how they evolve over time to be augmented with provenance of the workflow results and semantic annotations (Section 4).

2 Related Work

In certain disciplines (e.g., life sciences), scientific communication channels like journals encourage or mandate authors of submitted papers to include information about the methods used to reach the conclusions claimed in the paper. This has the aim of promoting reproducibility and reuse of the scientific results reported on those papers. For example, most 'wet lab' life science journal papers must contain a 'materials and methods' section that describes the details about the experiments that the authors conducted. These journals typically have strict rules about how to formulate these sections, but from a computational point of view it is weakly structured; hence they are still hard for other scientists to discover and reuse.

The practice of conveying computational methods in a standardised and highly structured way has had less time to evolve in many areas of science. Some journals are also encouraging authors to make available the data and software that have been used and produced, that is, to make data and processes used

part of the published work [8]. For example, Bioinformatics[1] considers software availability as an important prerequisite to the acceptance of the paper. And the NASA ADS (Astrophysics Data System)[2] is linking and referencing papers, references to the journal, data behind the plots used in the papers, catalogues of objects used (as URL references), software used (as URL references to the Astrophysics Source Code Library), instrument used to gather the observed/input data, and the proposal submitted to ask for observation time. These are important steps forward to promote sharing and reuse. However, software and data availability may not be sufficient to check the reproducibility of results, as described in the introduction.

As stated in the introduction, our model is built on earlier work on myExperiment packs [15], which aggregate elements such as workflows, documents and datasets together, following Web 2.0 and Linked Data principles [18, 17]. The myExperiment ontology [14], which forms the basis for our research object model, has been designed such that it can be easily aligned with existing ontologies. For instance, their elements can be assigned annotations comparable to those defined by Open Annotation Collaboration (OAC).

One important aspect of our work is that we make use of abstract workflow templates as a means to annotate workflow templates, facilitating workflow specification (as done by Gil *et al.* [6] and Ludascher *et al.* [9]). Scientists describe a workflow by identifying abstract tasks and specifying scientific analyses using semantic concepts from an underlying domain ontology. The specified abstract workflow is then mapped to a concrete workflow using mappings that specify for each task the underlying service operations that can be used for its implementations.

Our work is complementary to the above proposals in the sense that, in addition to semantic annotations of workflows, we exploit provenance of workflow results to describe workflow templates. In this context, similar proposals are CrowdLab [10], which provides users with the means for publishing data as well as workflows and the provenance of their results to promote the reproducibility of such results, Janus [12] and OPMW [5]. Here we leverage semantic technologies and underline the importance of annotations, which we hope will yield a wide adoption of research objects among scientists. Besides, we allow connecting more elements to the workflow: alternative material, alternative web services, bibliography, the proposal that led to the workflow/experiment, etc.

A clear demand from domains such as bioinformatics and astronomy is the ability to understand a workflow, for which elements outside of the workflow are often needed.

3 A Model for Workflow-Centric Research Objects

Our workflow-centric research object model aims at providing support for the description of the scientific processes described in the previous section in a machine

[1] http://bioinformatics.oxfordjournals.org/
[2] http://labs.adsabs.harvard.edu/

processable format, together with the datasets involved, the results obtained, and their provenance information. The research object will be also accompanied with annotations, which will promote the discover-ability, and therefore the reusability of the processes (workflows), as well as enabling third parties to assess the validity and reproducibility of the results.

Figure 1 illustrates a coarse-grained view of a workflow-centric research object, which aggregates a number of resources, namely:

- a workflow template, which defines the workflow;
- workflow runs obtained by enacting the workflow template
- other artifacts which can be of different kinds, e.g., a paper that describes the research, datasets used in the experiments, etc.;
- annotations describing the aforementioned elements and their relationships.

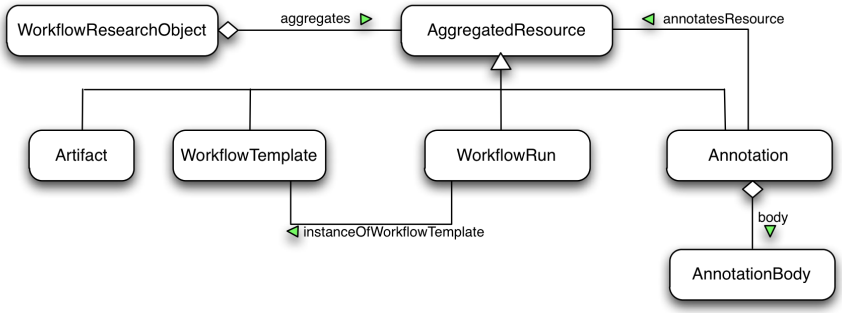


Fig. 1: Workflow-centric research object as an aggregation of resources. CHECK IF IT IS OK

Figure 2 provides a more detailed view of the resources that compose workflow templates and workflow runs. A workflow template is a graph in which the nodes are processes and the edges represent data links that connect the output of a given process to the input of another process, specifying that the artifacts produced by the former are used to feed the latter. A process is used to describe a class of actions that when enacted give rise to process runs. The process specifies the software component (e.g., web service) responsible for undertaking the action. Note that some workflow systems may specify in addition to the data flow, the control flow, which specifies temporal dependencies and conditional flows between processes. We chose to confine the workflow research object model to data-driven workflows, as in Taverna [16], Triana [2], the process run Network Director supplied by Kepler [4], Galaxy[3], Wings [7], etc.

Figure 3-b illustrates an example of a **workflow template** that is composed of two processes. Such a workflow describes an in-silico bioinformatics experiment that is used to identify gene pathways. Specifically, the workflow is composed of two processes: given a protein accession, the *GetKeggGeneId* process is used to

[3] http://galaxy.psu.edu/

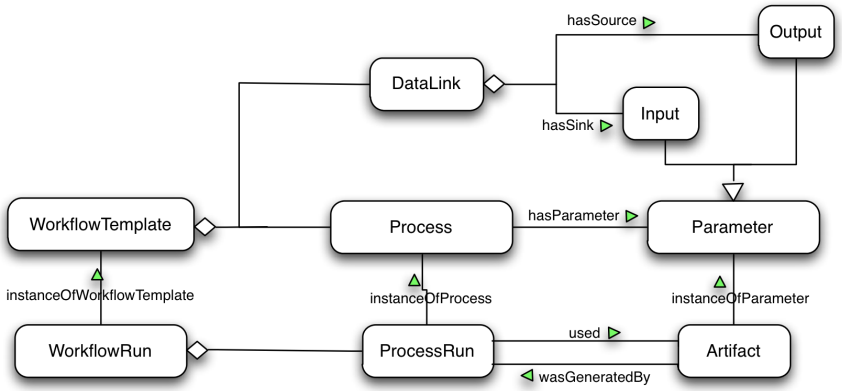


Fig. 2: Resources aggregated within workflow-centric research objects and their relationships. CHECK IF IT IS OK. for instance, abstract workflow does not appear, although it appears on the text, the same on the next figure.

retrieve the corresponding gene ID. The gene ID retrieved is then used to feed the *GetKeggPathway* process, which returns the corresponding pathways. Note that we also support workflow instances, which are workflow templates with the inputs bound to data values. We also distinguish between standard input parameters and configuration input parameters. Configuration input parameters are used to set the algorithm, the underlying sources used by the processes that compose a workflow template and so on. In addition, the processes that compose a workflow template are not always bound to a software component, rather they can be performed manually in which case they are associated with a human agent.

A workflow template can be instantiated and enacted using a workflow engine, e.g., Taverna. This gives rise to a workflow run that specifies the process runs that were obtained by executing the processes that constitute the workflow template in question. For example, when the action specified by the process is undertaken by a web service, the process run obtained by enacting such a process represents a web service call. A process run may take as input some existing artifacts, specified by the *used* association, and output some new artifacts, specified by the *wasGeneratedBy* association. Artifact is a general concept that represents an immutable piece of state, which may have a physical embodiment in a physical object, or a digital representation in a computer system [13]. In the context of workflow-centric research objects, the focus is on artifacts that are digital representations in a computer system. It is worth mentioning that the notion of process run and artifact that we use are aligned with major provenance models such as the Open Provenance Model (OPM) [13] and PROV-DM[4].

Figure 3-c illustrates an example of a **workflow run** that is obtained by enacting the workflow template together with the provenance of the results produced by the workflow run, which are depicted in Figure 3-b. *Get-*

[4] http://www.w3.org/TR/2011/WD-prov-dm-20111018

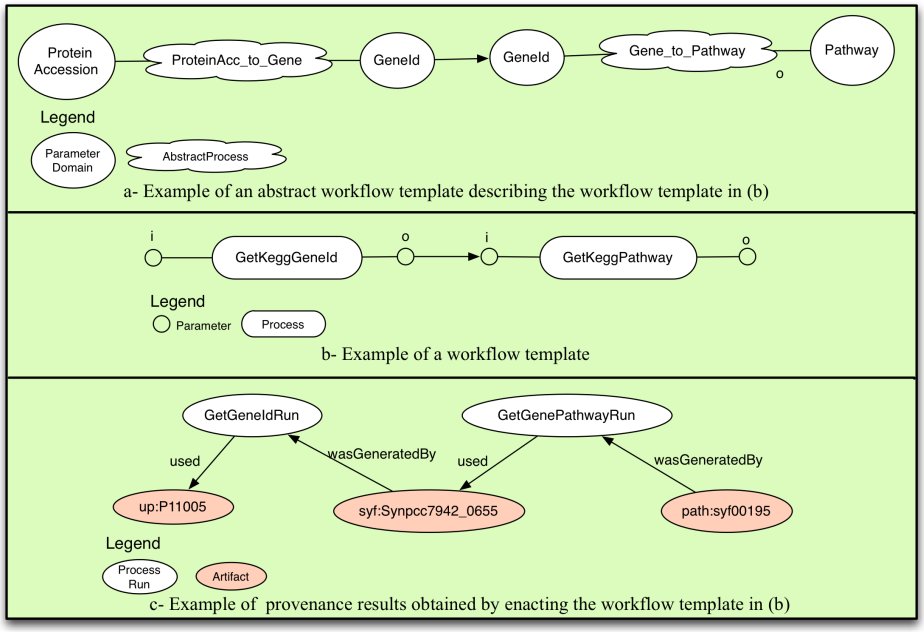


Fig. 3: Example of a workflow template (b), an abstract workflow (a) that semantically describes such workflow template, and provenance of workflow results (c) obtained by enacting the workflow template.

GeneIdRun, and *GetGenePathwayRun* are process runs that were obtained by enacting the *GetGeneId* and *GetGenePathway* processes, respectively. *GetGeneIdRun* took as input the protein accession *up:11005* and generated the gene id *syf:Synpcc7942_0655*, the process run *GetGenePathwayRun* then used *syf:Synpcc7942_0655* to generate the pathway *path:syf00195*.

It is important to highlight that scientists can annotate the elements of a workflow-centric research object (along with the research object itself). They can specify the title of a research object, its purpose, its version, ownership, citations, etc. A more accurate form of annotation can be used to describe the elements of a research object by linking them to concepts from domain ontologies. In particular, this kind of annotation can be used to effectively browse and query workflow templates.

Finally, workflow templates can be annotated in an **abstract workflow template**, which is a graph of abstract processes that are connected by data links. The abstract processes and their input and output parameters are labeled with concepts from underlying domain ontologies, e.g., [21, 22], which specify the tasks performed by the steps and the semantic domains of their parameters, respectively. An abstract worklfow template *awf*, which is used to annotate a given workflow template *wf*, has the same data flow topology as *wf*. The abstract processes that compose *awf* annotate the processes in *wf*, and the parameter domains in *awf* specify the semantic domains of the process parameters in *wf*. As

an example, Figure 3-a illustrates an abstract workflow template that semantically describes the workflow template depicted in Figure 3-b. *ProteinAcc_to_Gene* and *Gene_to_Pathway* are two concepts that specify the tasks of the processes *GetKeggGeneId* and *GetKeggPathway*, respectively, whereas *ProteinAccession*, *GeneId* and *Pathway* are concepts that specify the domain of the input and output parameters of such processes.

3.1 Grounding Workflow-centric Research Objects Using Semantic Technologies

Workflow-centric research objects are encoded using RDF[5], according to a set of ontologies that we have made available[6].

Following myExperiment packs, research objects use the Object Exchange and Reuse (ORE) model[7], to represent aggregation. ORE defines standards for the description and exchange of aggregations of Web resources. Using ORE, a workflow-centric research object is defined as a resource that aggregates other resources, i.e., workflow(s), provenance, other objects and annotations. For example, the RDF turtle snippet illustrated below specifies that a research object identified by `:wro` aggregates a workflow template `:pathway_wf_sp`, a workflow run `:pathway_wf_run`, and an annotation `:wf_annot`.

Example of a research object defined as an ORE aggregation

```
:wro a :WorkflowResearchObject , ore:Aggregation ;
     ore:aggregates  :pathway_wf_sp ,
                     :pathway_wf_run ,
                     :wf_annot .
:pathway_wf_sp  a :WorkflowTemplate .
:pathway_wf_run a :WorkflowRun .
:wf_annot a ao:Annotation .
```

We also use the Annotation Ontology (AO)[8], which provides a common model for annotating resources. This differs from myExperiment packs, which use a vocabulary that is mapped to Open Annotation Collaboration (OAC)[9][10]. Several types of annotations are supported by the Annotation Ontology, e.g., comments, textual annotations (classic tags) and semantic annotations which relate elements of the research objects to concepts from underlying domain ontologies. As an example, the RDF turtle snippet below shows how the abstract workflow template illustrated in Figure 3-a can be specified using a named graph `:pathway_abs_wf_graph`. It also shows how, using Annotation Ontology, such

[5] http://www.w3.org/RDF

[6] http://www.wf4ever-project.org/wiki/display/docs/Research+Object+Vocabulary+Specification

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

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

[9] www.openannotation.org

[10] Note that work is currently underway to align the two annotation vocabularies: http://www.w3.org/community/openannotation/

an abstract workflow template can be used to annotate the workflow template `:pathway_wf_sp`, which is depicted in Figure 3-b. Specifically, a resource representing the annotation, `:wf_annot`, is created to link the workflow template which is subject to annotation, `:pathway_wf_sp`, to the named graph specifying the corresponding abstract workflow template, `:pathway_abs_wf_graph`.

Example illustrating how a workflow template can be annotated using AO

```
:wf_annot a   ao:Annotation ;
          ao:annotatesResource :pathway_wf_sp ;
          ao:body :pathway_abs_wf_graph.
:pathway_abs_wf_graph {
        :pathway_wf_sp :hasAbsWorkflowTemplate :pathway_abs_wf.
        :pathway_abs_wf  a :AbsWorkflowTemplate ;
                 :hasAbsProcess :ap1,
                                :ap2 .
                 :hasDataLink :dl .
        :ap1 :hasTask :t1 ;
             :hasInput :ap1_in ;
             :hasOutput :ap1_out .
        :t1 a mygrid:ProteinAcc_to_Gene .
        :ap2 :hasTask :t2 ;
             :hasInput :ap2_in ;
             :hasOutput :ap2_out .
        :t2 a mygrid:Gene_to_Pathway .
        :ap1_in  :hasDomain :d1 .
        :ap1_out :hasDomain :d2 .
        :ap2_in  :hasDomain :d3 .
        :ap2_out :hasDomain :d4 .
        :d1 a  mygrid:ProteinAccession .
        :d2 a  mygrid:GeneId .
        :d3 a  mygrid:GeneId .
        :d4 a  mygrid:Pathway .
        :dl  :from :ap1_out ;
             :to :ap2_in . }
```

4 The Lifecycle of a Workflow-Centric Research Object

We will now illustrate research object lifecycle through a small example that shows how all the resources contained in a research object are bundled as the scientific experiment progresses. This example lifecycle is summarized graphically in Figure 4.

A research object normally starts its life as an empty **Live Research Object**, with a first design of the experiments to be performed (which determines what workflows and resources will be added, by either retrieving them from an existing platform or creating them from scratch). Then the research object is filled incrementally by aggregating such workflows that are being created, reused or re-purposed, datasets, documents, etc. Any of these components can be changed at any point in time, removed, etc.

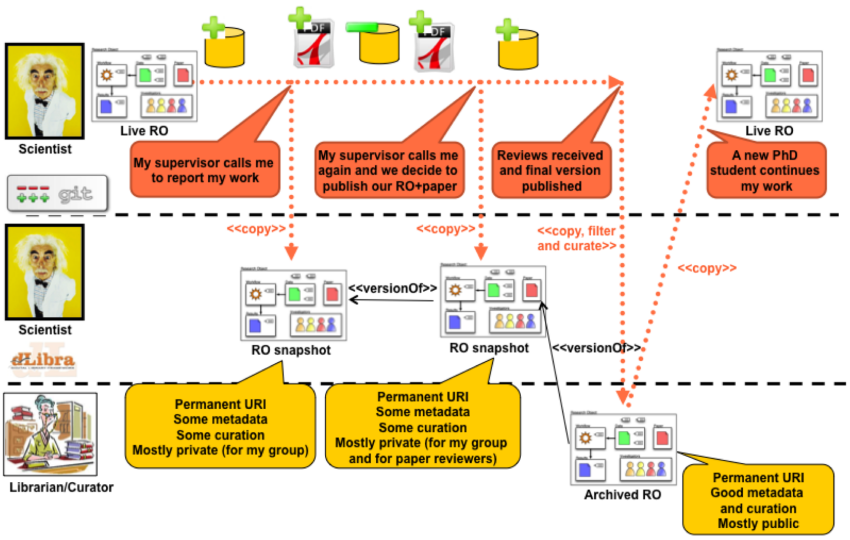


Fig. 4: A sample research object lifecycle.

In our scenario, we observe several points in time when this **Live Research Object** gets copied and kept into a **Research Object snapshot**, which aims to reflect the status of the research object at a given point in time. Such a snapshot may be useful to release the current version of the research outcome of an experiment, submit it to be peer reviewed or to be published (with the appropriate access control mechanisms), share it with supervisors or collaborators, or for acknowledgement and citation purposes.

A snapshot may also contain a paper describing the research object in general and the experiment in particular, depending on the policies of the corresponding scientific communication channel, e.g., workshop, conference or journal. Such snapshots have their own identifiers, and may even be preserved, since it may be useful to be able to track the evolution of the research object over time, so as to allow, for example, retrieval of a previous state of the research object, reporting to funding agencies the evolution of the research conducted, etc.

At some point in time, the research object may get published and archived, in what we know as an **Archived Research Object**, with a permanent identifier. Such a version of our research object may be the result of copying completely our **Live Research Object**, or it may be the result of some filtering or curation process where only some parts of the information available in the aggregation are actually published for others to reuse. As illustrated in Figure 4, a user can use an existing **Archived Research Object** as a starting point to his or her research, e.g., to repurpose it or its parts, in which case a new **Live Research Object** is created based on the existing **Archived Research Object**.

This is only one of the many potential scenarios that could be foreseen for the lifecycle of a workflow-centric research object and we are currently defining different storyboards for their evolution. One important aspect to highlight is

the fact that during its whole lifecycle, the research object is aggregating new objects. The annotation process during the lifecycle of experimentation allows the generation of sufficient metadata about the research objects to support preservation and sharing. Therefore, when a scientists decides to preserve it most of the annotations that will be needed for that preservation process will be already available inside the research object.

5 Conclusions and Further Work

Scientific workflows are used by scientists not only as computational units that encode scientific methods that can be shared among scientists, but also to specify their experiments. In this paper we presented a research object model to capture all the needed information and data including the methods (workflows) and other elements: namely annotations, datasets, provenance of the workflow results, etc.

We showed how this model has been implemented using semantic technologies reusing existing vocabularies, so that scientists are now able to query and publish their experiments according to existing standards. As a result, experiments may be more interoperable, since they are recorded with the same general model to describe them; they can be reused more easily; and decay can be better handled by representing the information of the templates and the traces in an environment/execution independent manner.

The work reported in this paper is preliminary. Our ongoing work includes the design of an architecture for the management of workflow-centric research objects, based on the model presented in this paper, which is being implemented and made available in the Wf4Ever sandbox (http://sandbox.wf4ever-project.eu/). We are also currently validating the model presented in this paper by creating research objects for existing workflows that are stored within the myExperiment repository. In doing so, we are examining issues that have to do with the decay of workflow, mechanisms for querying research objects, and scalability. As well as the technical challenges, we are aware that there are social challenges that need to be overcome to encourage scientists to adopt research object as a unit for publication, discovery and reuse of scientific communications. In this respect, we started collaborating with scientists from the European projects BioVeL (Biodiversity Virtual e-Laboratory)[11] and SCAPE (SCAlable Preservation Environments[12]).

Acknowledgements

The research reported in this paper is supported by the Wf4Ever project (http://www.wf4ever-project.org), Project 270129 funded under EU FP7 Digital Libraries and Digital Preservation (ICT-2009.4.1).

[11] http://www.biovel.eu
[12] http://www.scape-project.eu

References

1. S. Bechhofer, I. Buchan, D. D. Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch, D. Cruickshank, and Et Al. Why linked data is not enough for scientists. *Future Generation Computer Systems*, 2011.
2. David Churches and Et Al. Programming scientific and distributed workflow with triana services. *Concurrency and Computation: Practice and Experience*, 18(10):1021–1037, 2006.
3. Ewa Deelman and Et Al. Workflows and e-science: An overview of workflow system features and capabilities. *FGCS*, 25(5):528–540, 2009.
4. Lei Dou and Et Al. Scientific workflow design 2.0: Demonstrating streaming data collections in kepler. In *ICDE*, pages 1296–1299. IEEE Computer Society, 2011.
5. Daniel Garijo and Yolanda Gil. A new approach for publishing workflows: Abstractions, standards, and linked data. In *Proceedings of the Sixth Workshop on Workflows in Support of Large-Scale Science (WORKS'11), held in conjunction with SC 2011*, Seattle, Washington, 2011.
6. Yolanda Gil and Et Al. Mind your metadata: Exploiting semantics for configuration, adaptation, and provenance in scientific workflows. In *International Semantic Web Conference (2)*, pages 65–80. Springer, 2011.
7. Yolanda Gil, Varun Ratnakar, Jihie Kim, Pedro Antonio Gonzalez-Calero, Paul Groth, Joshua Moody, and Ewa Deelman. Wings: Intelligent workflow-based design of computational experiments. *IEEE Intelligent Systems*, 26(1), 2011.
8. Darrel C. Ince, Leslie Hatton, and John Graham-Cumming. The case for open computer programs. *Nature*, 482(7386):485–488, 02 2012.
9. Bertram Ludäscher, Ilkay Altintas, and Amarnath Gupta. Compiling abstract scientific workflows into web service workflows. In *SSDBM*, pages 251–254. IEEE Computer Society, 2003.
10. Phillip Mates, Emanuele Santos, Juliana Freire, and Cláudio T. Silva. Crowdlabs: Social analysis and visualization for the sciences. In *SSDBM*, pages 555–564. Springer, 2011.
11. Jill P. Mesirov. Accessible reproducible research. *Science*, 327(5964):415–416, 2010.
12. Paolo Missier, Satya S Sahoo, Jun Zhao, Carole Goble, and Amit Sheth. Janus: from workflows to semantic provenance and linked open data. *Life Sciences*, 6378(i):129–141, 2010.
13. Luc Moreau and Et Al. The open provenance model core specification (v1.1). *Future Generation Comp. Syst.*, 27(6):743–756, 2011.
14. David Newman. *The Building and Application of a Semantic Platform for an e-Research Society*. PhD thesis, UNIVERSITY OF SOUTHAMPTON, 2011. Submitted on October 2011.
15. David Newman, Sean bechhofer, and David De Roure. myexperiment: An ontology for e-research. In *Workshop on Semantic Web Applications in Scientific Discourse in conjunction with the International Semantic Web Conference*, 2009.
16. Thomas M. Oinn and Et Al. Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience*, 18(10):1067–1100, 2006.
17. Kevin R. Page, David De Roure, and Et Al. Rest and linked data: a match made for domain driven development? In *2nd International Workshop on RESTful Design (WS-REST 2011) held in conjunction with WWW 2011*, 2011.
18. David De Roure and Et Al. The evolution of myexperiment. In *e-Science 2010*. IEEE, 2010.

19. David De Roure, Sean Bechhofer, Carole A. Goble, and David R. Newman. Scientific social objects: The social objects and multidimensional network of the myexperiment website. In *SocialCom/PASSAT*. IEEE, 2011.
20. David De Roure, Khalid Belhajjame, and Et Al. Towards the preservation of scientific workflows. In *Procs. of the 8th International Conference on Preservation of Digital Objects (iPRES 2011)*. ACM, 2011.
21. Vuong Xuan Tran and Hidekazu Tsuji. Owl-t: A task ontology language for automatic service composition. In *ICWS*, pages 1164–1167. IEEE Computer Society, 2007.
22. Chris Wroe, Robert Stevens, Carole A. Goble, Angus Roberts, and R. Mark Greenwood. A suite of daml+oil ontologies to describe bioinformatics web services and data. *Int. J. Cooperative Inf. Syst.*, 12(2):197–224, 2003.

Using annotations to model discourse: an extension to the Annotation Ontology

Leyla Jael García-Castro[1], Olga Giraldo[2], Alexander García[3]

[1] Universität der Bundeswehr München, Werner-Heisenberg-Weg 39,
85779 Neubiberg, Germany
w31blega@unibw.de
[2] Universidad Politécnica de Madrid, Ontology Engineering Group,
Madrid, Spain
oxgiraldo@gmail.com
[3] Florida State University,
Tallahassee, Florida, USA
alexgarciac@gmail.com

Abstract. The Annotation Ontology (AO) has proven to be a valuable resource for structuring annotations in scientific documents. We are representing elements of discourse with the AO; by using our proposed extension it is possible to mark up specific rhetorical structures and build a network of interconnected documents. The extension presented in this paper also makes it possible to represent more expressive associations across nanopublications.

Keywords: Scientific publications, social tagging systems, social and semantic web, knowledge discovery

1 Introduction

Digital Libraries such as Elsevier Science Direct[1] or PubMed[2] store electronic versions of scientific publications. Although these resources provide some information retrieval mechanisms, it is still difficult to extract facts buried in the text [1]; for instance, retrieving definitions and claims from literature is usually a manual process. Making the content explicitly identifiable by means of Semantic Web (SW) technology has been proposed as a feasible solution for improving information retrieval across digital libraries; enriching the metadata should make it possible to identify and extract facts buried in documents [1, 2]. Documents should be self-descriptive and fully immersed in the web of data [3].

Heading towards a self-descriptive document requires a well-organized annotation structure consistent with the underlying rhetorical structure. Annotation should support not only marking segments but also making the relationships across these portions explicit. Furthermore, annotations should scaffold relations across documents. It is not enough to know the concepts in a document; it is also necessary

[1] http://www.sciencedirect.com/
[2] http://www.ncbi.nlm.nih.gov/pubmed/

to know how are they related [4] -within the document, across documents and to the web of data.

The Annotation Ontology (AO) [5] facilitates modeling annotations on static resources; an ongoing project will extend the AO in order to facilitate the annotation on mutable objects as well [6]. The AO is built upon the Annotea Project[3] and supports both free and semantic annotations: free annotations are expressed by plain text attached to resources whilst semantic annotations should also include a relation *ao:hasTopic* to an ontological entity. Annotations can be attached to the whole resource but also to portions of it, *e.g.* sentences, paragraphs, sections, images, tables, etc. Annotations on any fragment within a document should be modeled by using selectors; a selector identifies the fragment depending on its nature: *aos:TextSelector* identifies a exact match to a piece of text, *aos:StartEndSelector* identifies the initial and final position that the annotation refers to, *aos:InitEndCornerSelection* identifies the initial and final (x,y) coordinates within an image, etc. The AO offers several types of annotations such as notes, comments, erratum, etc. Qualifiers are a particular type of annotations mapped to the Simple Knowledge Organization System[4] (SKOS) properties and particularly useful for semantic annotations: *ao:Qualifer* maps to *skos:RelatedMatch*, *ao:ExactQualifer* to *skos:exactMatch*, *ao:CloseQualifier* to *skos:closeMatch*, *ao:BroadQualifer* to *skos:broadMatch*, and *ao:NarrowQualifier* to *skos:narrowMatch*. The provenance within AO is supported by the Provenance Authoring and Versioning ontology[5] that provides features on provenance to support scientific content and its curation. Scientific discourses are modeled by integrating the AO and SWAN [7].

We are broadening the interoperability between SWAN and AO [5], going beyond the current integration[6]. We are including concepts from CoreSC [8], SWAN, the Sample Processing and Separation Techniques (SEP), Ontology for Biomedical Investigations (OBI), Micro Array Gene Expression Ontology (MGED), and National Cancer Institute Thesaurus (NCIt). On the one hand, we want to make explicit some discursive elements in scientific publications, for instance, the structural elements related to the arrangement and distribution of the document. We are also interested in identifying argumentative elements, *i.e.* elements of the discourse. On the other hand, we are adding new qualifiers and representing annotations on annotations to express relations across elements and the initiation of a topic thread, *i.e.* an argumentative line anchored in the document. We have initially focused our efforts in modeling literature reviews; although these papers summarize findings reported in other documents and offer insightful analysis of existing literature, extracting claims, definitions, data, and other data types is cumbersome –partly due to the lack of markers for these structures. Moreover, as literature reviews bring together information from existing documents by pulling out facts and structuring them in a new document, such a network is not explicit; we are providing the structure so that literature reviews can be seen as a collection of scaffolded annotations and/or nanopublications.

[3] http:// www.w3.org/2001/Annotea/
[4] http:// www.w3.org/2004/02/skos/
[5] PAV, swan.mindinformatics.org/spec/1.2/pav.html
[6] http://code.google.com/p/annotation-ontology/wiki/SWANDiscourse

2 Rhetoric and discourse from Annotations

2.1 AO extension to model rhetoric and discourse elements

We have extended the AO with new classes and properties that facilitate making explicit the rhetoric and discourse embedded in a scientific publication. Classes are meant to categorize the type of structures expressed in a publication -*e.g.* definitions, examples, claims, etc. From the AO, we have reused *ao:Definition* making it compliant to the Meaning-of-a-tag (MOAT) [9]. We also reused *ao:Example* as it was originally proposed. In addition, we integrated classes from SWAN vr. 1.2[7] as well as from CoreSC. Table 1 summarizes the proposed classes and presents a short description of the intended purpose. Whenever a class matches an entity from another vocabulary the description is taken from there, descriptions taken from the Cambridge Dictionaries Online[8] are identified as CDO, and no quoted descriptions are defined by the authors.

Table 1. Classes modeling concepts in a scientific publication

Class name	Description
aold:Introduction	Section used to broadly present the problem, existing solutions, and what the research work intends to achieve
aold:Motivation	coresc:Motivation, "The reasons behind an investigation"
aold:Aim	CDO "a result that your plans or actions are intended to achieve"
aold:Goal	coresc:Goal, "A target state of the investigation where intended discoveries are made"
aold:Hypothesis	coresc:Hypothesis, "A statement not yet confirmed rather than a factual statement"
ao:ResearchQuestion	swan:ResearchQuestion
ao:ResearchStatement	swan:ResearchStatement
aold:Reference	coresc:Background, "Generally accepted background knowledge and previous work"
aold:Report	obi:report "a document assembled by an author for the purpose of providing information for the audience. A report is the output of a documenting process and has the objective to be consumed by a specific audience."
aold:Counter-Example	Example that contradicts a statement or idea
aold:Opinion	CDO "a thought or belief about something or someone, a judgment about someone or something"
aold:Claim	CDO "to say that something is true or is a fact, although you cannot prove it and other people might

[7] http://swan.mindinformatics.org/ontologies/1.2/discourseelements.owl
[8] http://dictionary.cambridge.org/

	not believe it"
aold:Method	NCIt "A means, manner of procedure, or systematic course of actions that have to be performed in order to accomplish a particular goal."
aold:Sample	SEP "A sample is a substance role played by a biological substance as an input substance to a protocol."
aold:Protocol	OBI "a protocol is a plan specification which has sufficient level of detail and quantitative information to communicate it between domain experts, so that different domain experts will reliably be able to independently reproduce the process."
aold:Model	coresc:Model, "A statement about a theoretical model or framework"
aold:Experiment	MGED "The complete set of assays and their descriptions performed as an experiment for a common purpose."
aold:ObservationInResearch	NCIt "Watching something and taking note of what happens."
aold:Result	coresc:Result, "Factual statements about the outputs of an investigation"
aold:Discussion	The annotation identifies a fragment corresponding to the discussion of the document
aold:Conclusion	coresc:Conclusion, "Statements inferred from observations & results relating to research hypothesis"

We are also proposing classes that facilitate the definition of relations between entities; this makes it possible to relate fragments within the same document or across multiple documents. Qualifiers in the AO are mapped to SKOS properties; here we interpret them as expressing a subjacent relationship between the annotated fragment/document and the topic. It is recommended to use ao:Annotation+*ao:hasTopic* whenever it is needed to point to examples, definitions and external links related to URIs. The ao:Qualifier/***aold:OnFlyQualifier***+*ao:hasTopic* should be used to relate the annotated fragment/document to an entity/resource, *e.g.* to relate "mouse" to an ontological term "ncbitaxon:10090". Qualifiers, as proposed by AO, express only five relationships, we are proposing ***aold:OnFlyQualifier*** that extends the original *ao:Annotation* to model any relationship that has been defined somewhere else, typically an ontology. The property ***aold:definesRelation*** maps to the subjacent relation, *e.g. owl:sameAs*, between the annotated fragment/document and the topic. Relations used by a specific annotation project can be narrowed down to a set of predefined relations; they may also be open to represent any relation expressed as free text. The last scenario would probably require curation mechanisms to see whether the relation is new or can be mapped to an existing one. Fig. 1 shows two possible uses of ***aold:OnFlyQualifier****:* on the left, an annotator has identified the paper with URI "http://biotea.ws/paper1" as the same at "http://tinyurl.com/apaper"; on the right, an annotator has identified a claim and its corresponding source.

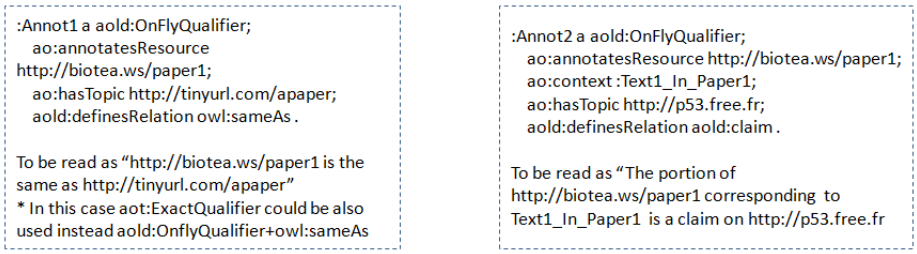


Fig. 1. aold:OnFlyQualifier in use

The ***aold:Relation*** extends the ***aold:OnFlyQualifier***, it represents an annotation that brings together a pair of annotations; the *ao:body* of the annotation establishes the intended name for the annotation. In this way, it is possible to use ***aold:OnFlyQualifier*** for known relations while using the ***aold:Relation*** may be reserved for new ones. Two subclasses have been proposed, ***aold:UnidirectionalRelation*** for those relations where the subject and object cannot be interchanged, *e.g. is_a*, and ***aold:BidirectionalRelation*** for all other relations, e.g. synonyms. If ***aold:definesRelation*** is used, it defines a relation, e.*g. sameAs*, between the *ao:body* and the topic. Fig. 2 shows a graphical representation of these classes.

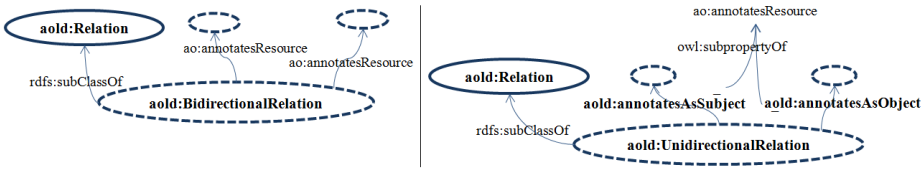


Fig. 2. aold:Relation, unidirectional and bidirectional

aold:Relation is useful in cases such as definitions in vocabularies as well as when making explicit a process described by a document. Additional information can be found at http://biotea.ws. Fig. 3 illustrates how could ***aold:Relation*** be used when expressing new ways to relate documents. An annotated fragment in a document is categorized as an "opinion" on a fragment of a second document.

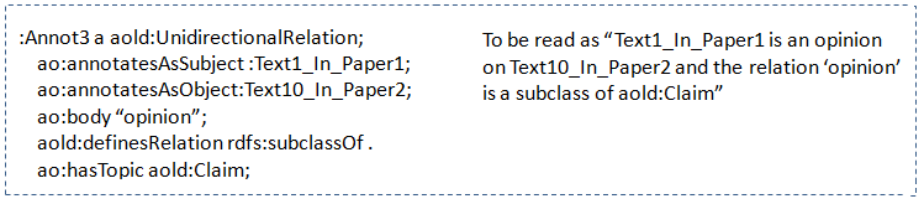


Fig. 3. aold:UnidirectionalRelation

We have also added two new selectors that work on RDF documents. When working with these selectors it is assumed that only one ***rdfs:comment*** will be present

in the RDF element being annotated. The ***aold:ElementSelector*** extends the ***aos:textSelector,*** it identifies an exact chunk of text in the ***rdfs:comment*** while the ***aold:StarEndElementSelector*** extends the **ao:StartEndSelector** by identifying the *start* and *end* positions of the text being annotated in the ***rdfs:comment***. Fig. 4 illustrates an example using the latter selector: A text from position 27 to position 38 in the "*Introduction*" section of the corresponding RDF representation for the paper with DOI 10.1016/SO014-5793(03)00051-6 has been annotated; the annotation body is "*mus musculus*". The ***aold:OnFlyQualifier*** is here used to indicate that the annotated text corresponds to *ncbitaxon:10090*.

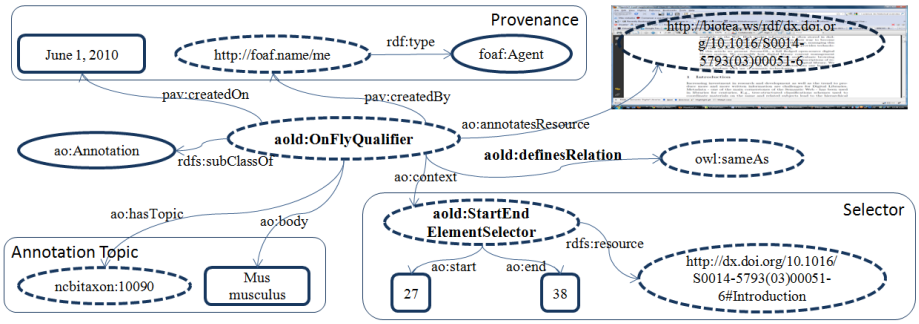


Fig. 4. aold:OnFlyQualifier and aold:StartEndElement selector in use

In order to support the argumentative process hidden in the text, we are reusing the properties proposed by the SWAN [7]; we are also adding new properties, some of them based on [4]. Table 2 summarizes these properties.

Table 2. Classes modeling concepts in a scientific publication

Property name	Description
aold:supportedBy	To specify where the support for the annotated fragment/document can be found (inverse *aold:supports*)
aold:contradicts	When the annotated fragment/document expresses an opposite idea (inverse aold:contradictedBy)
aold:takenFrom	When a fragment has been taken from another text not mentioned as a reference (inverse *aold:takenIn*)
aold:introducedBy	To specify a term, concept, definition introduced by a document
aold:proves	When the annotated fragment/document offers proof (inverse *aold:provedBy*)
aold:rebuts	When the annotated fragment/document offers a rebuttal (inverse *aold:rebutedBy*)
aold:useDataFrom	To specify a data source used in a document (inverse *aold:dataUsedAt*)
aold:cites	similar to bibo:cites but without restrictions on domain and range (inverse *aold:citedBy*, similar to bibo:citedBy)

Figures 5 and 6 present a hypothesis in a document using the definition given in a different document; Fig. 5 uses the ***aold:OnFlyQualifier*** to establish the relation "*cites to*" whereas Fig. 6 uses ***aold:UnidirectionalRelation***. Both figures illustrate how a fragment in the document with DOI 10.1016/SO014-5793(03)00051-6 has been annotated as a hypothesis; this hypothesis cites a fragment in the document PMC1435992 that corresponds to a definition that has been identified as the same concept defined by the entity CHEBI_16113.

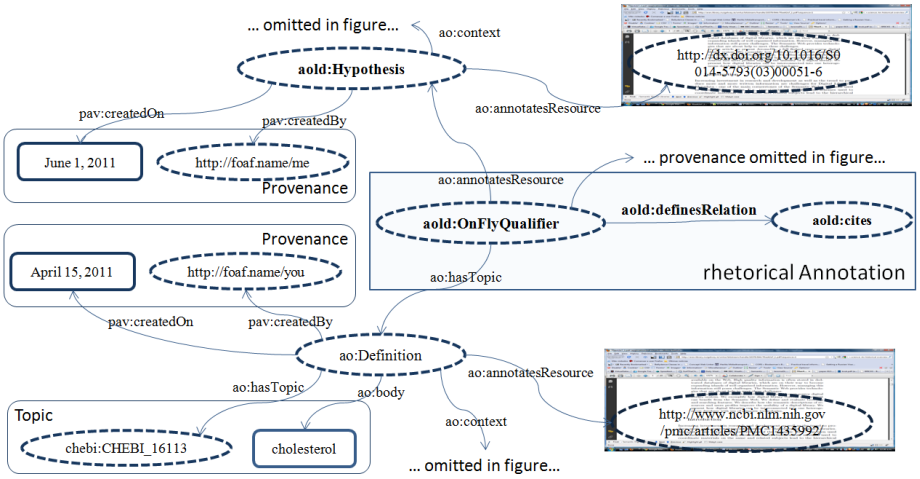


Fig. 5. aold:OnFlyQualifier - relating two documents using a known relation

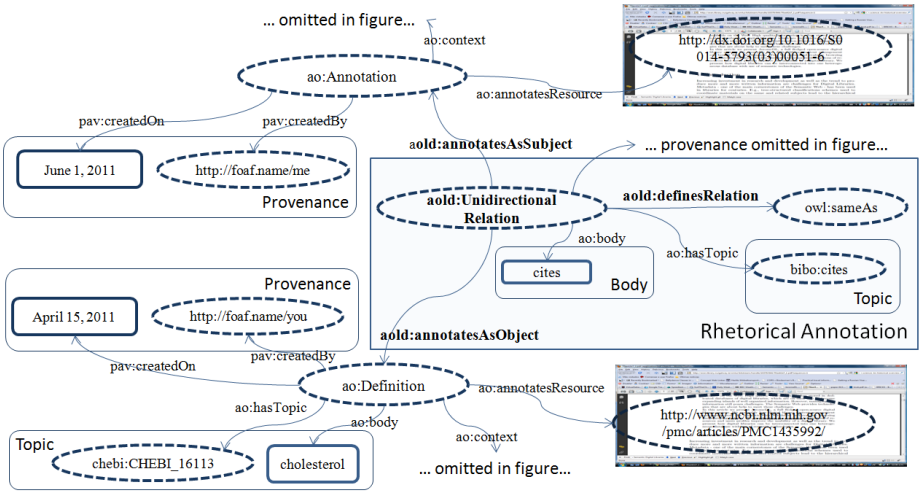


Fig. 6. aold:UnidirectionalRelation - relating two documents proposing a new relation

2.2 Nanopublications with the extended AO

Nanopublications are "a set of annotations that refers to the same statement and contains a minimum set of (community) agreed-upon annotations"; it is therefore feasible to use the proposed extension to represent nanopublications. Consider for instance the definitions for ontology; these vary depending on the field. One that is commonly used comes from Gruber, "*An ontology is a formal specification of a conceptualization*" [10]. The statement comprises three concepts: "an ontology", "is a", and "formal specification of a conceptualization". Upon this statement, different annotations are possible; for instance: (i) Tom Gruber is the author of the statement, (ii) the statement is a definition introduced at http://tomgruber.org/writing/ontolingua-kaj-1993.pdf, (iii) this statement is cited by a particular paper, and (iv) this statement is extended by a particular person. Fig. 7 shows these annotations; provenance of the annotations has been omitted on the sake of readability.

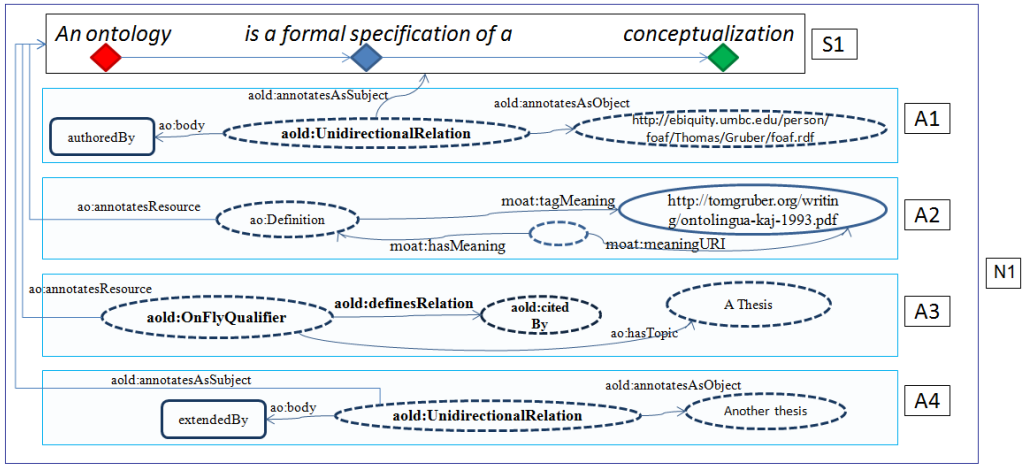


Fig. 7. A nanopublication

3 Discussion

The proposed extension to AO facilitates making explicit the rhetoric and discourse hidden in scientific publications in a way such that that machines can consume data. By using the extensions we can retrieve a list of publications related to a particular term, written by a specific author, or citing a specific gene or protein; for instance:

- All documents cited by document A that contains definitions coined in documents B, C, or D
- All materials used in documents cited by document A
- All documents from 2010 using method A but not method B
- All protocols used when materials A, B, and C have been also used

- All documents including some particular words (or entities) in a specific structural section, *e.g.* aim, thesis, discussion, results, etc.

Scientific publications annotated in the proposed way benefit from the Semantic Web and Link Open Data initiatives. Annotations and information extraction based on these publications become easier, as do sharing them and enriching them with other information also available in RDF. In this way, we facilitate the integration between literature and databases making it easier to use data available in publications for additional analysis. Our ultimate goal is to increase the pace of available scientific data by helping researchers to find information relevant to their projects. As publications are annotated, their rhetoric becomes searchable, thus researchers can better focus on those publications meeting their needs depending on what they are looking for, *e.g.* similar experiments based on methods, materials and protocols, or rebuttals of a particular theory. Furthermore, as annotations can be linked to databases, this enriched content can also be used for more specialized queries.

We have introduced our own description for some existing concepts in CoreSC such as *Experiment* and *Observation,* as we wanted them to be more accurate from the workflow laboratory perspective. We have not used the relations proposed by AZ-II [8] as they use the same category to express a relation and its corresponding inverse. For instance, *Support* is described as "Other work supports current work or is supported by current work".

Our approach is compatible with the principles of nanopublications. A concept would be a minimal *ao:Annotation* while relations on annotations could be used to define statements that are uniquely identified by a URI. Annotations, as they are understood in nanopublications, are also possible as relations, *i.e.* statement, are resources that can use as the subject of an annotation. The nanopublication itself becomes concrete by using the Annotation Set proposed by AO vr. 2.0; the Annotation Set is a container of annotations that is used to organize annotations that can be referred to as a whole. Furthermore, our approach makes it possible to relate nanopublications to any other type of publication. Our approach is also compatible with other annotation models, such as MOAT and Tag Ontology[9], making it easier to extend and integrate existing applications and tools.

4 Conclusions

We have presented an extension to AO that facilitates modeling rhetoric and discourse in scientific publications. Our approach entails using the common practice of annotating in order to identify specifics within the text; we are also gathering relationships between the annotated and referenced objects. Categories make it easier to identify whether it is about a claim, an example, a report, etc.; some of these categories come from the AO, others are new. In addition to the terms used during the exercise, we also worked with terms such as hypothesis, conclusion and research question; these are useful when analyzing the structure of scientific publications structure. Although not analyzed here, it could also be useful for commercial

[9] http://www.holygoat.co.uk/projects/tags/

documents, since they are also related to each other as well as to external resources. How to implement our model? This is a question beyond the scope of this paper that remains open; it is one of our top priorities.

Acknowledgments. We are grateful to Paolo Ciccarese for his fruitful discussions related to Hypertag migration to AO, *i.e.* using annotations to express relations.

References

1. Clare, A., Croset, S., Grabmüller, C., Kafkas, S., Liakata, M., Oellrich, A., Rebholz-Schuhmann, D.: Exploring the Generation and Integration of Publishable Scientific Facts Using the Concept of Nano-publications. SePublica - Workshop on Semantic Publishing, Vol. 721. CEUR, Heraklion, Crete, Greece (2011)
2. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Information Services and Use 30 (2010) 51-56
3. Garcia, A., Garcia, L.-J., Labarga, A., Giraldo, O., Montana, C., Bateman, J.: The Semantic Web and the Social Web heading towards a Living Document in life sciences. Journal of the Semantic Web. (2009)
4. de Waard, A., Breure, L., Kircz, J.G., Van Oosterndorp, H.: Modeling rhetoric in scientific publication. International Conference on Multidisciplinary Information Sciences and Technologies, Merida, Spain (2006) 1-5
5. Ciccarese, P., Ocana, M., Garcia Castro, L., Das, S., Clark, T.: An open annotation ontology for science on web 3.0. Journal of Biomedical Semantics 2 (2011) S4
6. Morris, R.A., Dou, L., Hanken, J., Kelly, M., Lowery, D., Ludaescher, B., Macklin, J.A., Morris, P.J.: Semantic Annotation of Mutable Data. To be submitted (2012)
7. Ciccarese, P., Wu, E., Wong, G., Ocana, M., Kinoshita, J., Ruttenberg, A., Clark, T.: The SWAN biomedical discourse ontology. Journal of Biomedical Informatics 41 (2008) 739-751
8. Liakata, M., Teufel, S., Siddharthan, A., Batchelor, C.: Corpora for the conceptualisation and zoning of scientific papers. International Conference on Language Resources and Evaluation, Malta (2010)
9. Passant, A., Laublet, P.: Meaning Of A Tag: A Collaborative Approach to Bridge the Gap Between Tagging and Linked Data. International World Wide Web Conference - Linked Data on the Web Workshop, China (2008)
10. Gruber, T.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 5(2) (1993) 199-220

Three Steps to Heaven: Semantic Publishing in a Real World Workflow

Phillip Lord, Simon Cockell, Robert Stevens

Newcastle and Manchester

Abstract. Semantic publishing offers the promise of computable papers, enriched visualisation and a realisation of the linked data ideal. In reality, however, the publication process contrives to prevent richer semantics while culminating in a 'lumpen' PDF. In this paper, we discuss a web-first approach to publication, and describe a three-tiered approach which integrates with the existing authoring tooling. Critically, although it adds limited semantics, it does provide value to all the participants in the process: the author, the reader and the machine.

1 Introduction

The publishing of both data and narratives on those data are changing radically. Linked Open Data and related semantic technologies allow for semantic publishing of data. We still need, however, to publish the narratives on that data and that style of publishing is in the process of change; one of those changes is the incorporation of semantics [1,2,3]. The idea of semantic publishing is an attractive one for those who wish to consume papers electronically; it should enhance the richness of the computational component of papers [2]. It promises a realisation of the vision of a next generation of the web, with papers becoming a critical part of a linked data environment [1,4], where the results and naratives become one.

The reality, however, is somewhat different. There are significant barriers to the acceptance of semantic publishing as a standard mechanism for academic publishing. The web was invented around 1990 as a light-weight mechanism for publication of documents. It has subsequently had a massive impact on society in general. It has, however, barely touched most scientific publishing; while most journals have a website, the publication process still revolves around the generation of papers, moving from Microsoft Word or LaTeX [5], through to a final PDF which looks, feels and is something designed to be printed onto paper[1]. Adding semantics into this environment is difficult or impossible; the content of the PDF has to be exposed and semantic content retro-fitted or, in all likelihood, a complex process of author and publisher interaction has to be devised and followed. If semantic data publishing and semantic publishing of academic narratives are to work together, then academic publishing needs to change.

[1] This includes conferences dedicated to the web and the use of web technologies.

In this paper, we describe our attempts to take a commodity publication environment, and modify it to bring in some of the formality required from academic publishing. We illustrate this with three exemplars - different kinds of knowledge that we wish to enhance. In the process, we add a small amount of semantics to the finished articles. Our key constraint is the desire to add value for all the human participants. Both authors and readers should see and recognise additional value, with the semantics a useful or necessary byproduct of the process, rather than the primary motivation. We characterise this process as our "three steps to heaven", namely:

- make life better for the machine to
- make life better for the author to
- make life better for the reader

While requiring additional value for all of these participants is hard, and places significant limitations on the level of semantics that can be achieved, we believe, it does increase the likelihood that content will be generated in the first place, and represents an attempt to enable semantic publishing in a real-world workflow.

2 Knowledgeblog

The knowledgeblog project stemmed from the desire for a book describing the many aspects of ontology development, from the underlying formal semantics, to the practical technology layer and, finally, through to the knowledge domain [6]. However, we have found the traditional book publishing process frustrating and unrewarding. While scientific authoring is difficult in its own right, our own experience suggests that the *publishing* process is extremely hard-work. This is particularly so for multi-author collected works which are often harder for the editor than writing a book "solo". Finally, the expense and hard copy nature of academic books means that, again in our experience, few people read them.

This contrasts starkly with the web-first publication process that has become known as blogging. With any of a number of ready made platforms, it is possible for authors with little or no technical skill, to publish content to the web with ease. For knowledgeblog ("kblog"), we have taken one blogging engine, WordPress [7], running on low-end hardware, and used it to develop a multi-author resource describing the use of ontologies in the life sciences (our main field of expertise). There are also kblogs on bioinformatics[2] and the Taverna workflow environment[3] [8]. We have previously described how we addressed some of the social aspects, including attribution, reviewing and immutablity of articles[6].

As well as delivering content, we are also using this framework to investigate *semantic academic publishing*, investigating how we can enhance the machine interpretability of the final paper, while living within the key constraint of making

[2] http://bioinformatics.knowledgeblog.org
[3] http://taverna.knowledgeblog.org

life (slightly) better for machine, author and reader without adding complexity for the human participants.

Scientific authors are relatively conservative. Most of them have well-established toolsets and workflows which they are relatively unwilling to change. For instance, within the kblog project, we have used workshops to start the process of content generation. For our initial meeting, we gave little guidance on authoring process to authors, as a result of which most attempted to use WordPress directly for authoring. The WordPress editing environment is, however, web-based, and was originally designed for editing short, non-technical articles. It appeared to not work well for most scientists.

The requirements that authors have for such 'scientific' articles are manifold. Many wish to be able to author while offline (particularly on trains or planes). Almost all scientific papers are multi-author, and some degree of collaboration is required. Many scientists in the life sciences wish to author in Word because grant bodies and journals often produce templates as Word documents. Many wish to use LaTeX, because its idiomatic approach to programming documents is unreplicable with anything else. Fortunately, it is possible to induce WordPress to accept content from many different authoring tools, including Word and LaTeX[6].

As a result, during the kblog project, we have seem many different workflows in use, often highly idiosyncratic in nature. These include:

Word/Email: Many authors write using MS Word and collaborate by emailing files around. This method has a low barrier to entry, but requires significant social processes to prevent conflicting versions, particularly as the number of authors increases.

Word/Dropbox: For the taverna kblog, authors wrote in Word and collaborated with Dropbox.[4] This method works reasonably well where many authors are involved; Dropbox detects conflicts, although cannot prevent or merge them.

Asciidoc/Dropbox: Used by the authors of this paper. Asciidoc[5] is relatively simple, somewhat programmable and accessible. Unlike LaTeX which can be induced to produce HTML with effort, asciidoc is designed to do so.

Of these three approaches probably the Word/Dropbox combination is the the most generally used.

From the readers perspective, a decision that we have made within knowledgeblog is to be "HTML-first". The initial reasons for this were entirely practical; supporting multiple toolsets is hard, particularly if any degree of consistency is to be maintained; the generation of the HTML is at least partly controlled by the middleware – WordPress in kblog's case. As well as enabling consistency of presentation it also, potentially, allows us to add additional knowledge; it makes semantic publication a possibility. However, we are aware that knowledgeblog currently scores rather badly on what we describe as the "bath-tub test"; while

[4] http://www.dropbox.com
[5] http://www.methods.co.nz/asciidoc/

exporting to PDF or printing out is possible, the presentation is not as "neat" as would be ideal. In this regard (and we hope only in this regard), the knowledge-blog experience is limited. However, increasingly, readers are happy and capable of interacting with material on the web, without print outs.

From this background and aim, we have drawn the following requirements:

1. The author can, as much as possible, remain within familiar authoring environments;
2. The representation of the published work should remain extensible to, for instance, semantic enhancements;
3. The author and reader should be able to have the amount of "formal" academic publishing they need;
4. Support for semantic publishing should be gradual and offer advantages for author and reader at all stages.

We describe how we have achieved this with three exemplars, two of which are relatively general in use, and one more specific to biology. In each case, we have taken a slightly different approach, but have fulfilled our primary aim of making life better for machine, author and reader.

3 Representing Mathematics

The representation of mathematics is a common need in academic literature. Mathematical notation has grown from a requirement for a syntax which is highly expressive and relatively easy to write. It presents specific challenges because of its complexity, the difficulty of authoring and the difficulty of rendering, away from the chalk board that is its natural home.

Support for mathematics has had a significant impact on academic publishing. It was, for example, the original motivation behind the development of TEX [9], and it still one of the main reasons why authors wish to use it or its derivatives. This is to such an extent that much mathematics rendering on the web is driven by a TEX engine somewhere in the process. So MediaWiki (and therefore Wikipedia), Drupal and, of course, WordPress follow this route. The latter provides plugin support for TEX markup using the `wp-latex` plugin [10]. Within kblog, we have developed a new plugin called `mathjax-latex` [11]. From the kblog author's perspective these two offer a similar interface – differences are, therefore, described later.

Authors write their mathematics directly as TEX using one of the four markup syntaxes. The most explicit (and therefore least likely to happen accidentally) is through the use of "shortcodes".[6] These are a HTML-like markup originating from some forum/bulletin board systems. In this form an equation would be entered as `[latex]e=mc^2[/latex]`, which would be rendered as "$e = mc^2$". It is also possible to three other syntaxes which are closer to math-mode in TEX: `$$e=mc^2$$`, `$latex e=mc^2$`, or `\[e=mc^2\]`.

[6] http://codex.wordpress.org/Shortcode

From the authorial perspective, we have added significant value, as it is possible to use a variety of syntaxes, which are independent of the authoring engine. For example, a TEX-loving mathematician working with a Word-using biologist can still set their equations using TEX syntax; although Word will not render these at authoring time but, in practice, this causes few problems for such authors, who are experienced at reading TEX. Within an LATEX workflow equations will be renderable both locally with source compiled to PDF, and published to WordPress.

There is also a W3C recommendation, MathML for the representation and presentation of mathematics. The kblog environment also supports this. In this case, the equivalent source appears as follows:

```
<math>
    <mrow>
        <mi>E</mi>
        <mo>=</mo>
        <mrow>
            <mi>m</mi>
            <msup>
                <mi>c</mi>
                <mn>2</mn>
            </msup>
        </mrow>
    </mrow>
</math>
```

One problem with the MathML representation is obvious: it is very long-winded. A second issue, however, is that it is hard to integrate with existing workflows; most of the publication workflows we have seen in use will on recognising an angle bracket turn it into the equivalent HTML entity. For some workflows (LATEX, asciidoc) it is *possible*, although not easy, to prevent this within the native syntax.

It is also possible to convert from Word's native OMML ("equation editor") XML representation to MathML, although this does not integrate with Word's native blog publication workflow. Ironically, it is because MathML shares an XML based syntax with the final presentation format (HTML) that the problem arises. The shortcode syntax, for example, passes straight-through most of the publication frameworks to be consumed by the middleware. From a pragmatic point of view, therefore, supporting shortcodes and TEX-like syntaxes has considerable advantages.

For the reader, the use of `mathjax-latex` has significant advantages. The default mechanism within WordPress uses a math-mode like syntax `$latex e=mc^2$`. This is rendered using a TEX engine into an image which is then incorporated and linked using normal HTML capabilities. This representation is opaque and non-semantic; it has significant limitations for the reader. The images are not scalable – zooming in cases severe pixalation; the background to the mathematics is coloured inside the image, so does not necessarily reflect the local style.

Kblog, however, uses the MathJax library[12]; this has a number of significant advantages for the reader. First, where the browser supports them, MathJax uses webfonts to render the images; these are scalable, attractive and standardized. Where they are not available, MathJax can fall-back to bitmapped fonts. The reader can also access additional functionality: clicking on an equation will raise a zoomed in popup; while the context menu allows access to a textual representation either as TEX or MathML irrespective of the form that the author used. This can be cut-and-paste for further use. Kblog uses the MathJax library[12] to render the underlying TEX directly on the client.

Our use of MathJax provides no significant disadvantages to the middleware layers. It is implemented in JavaScript and runs in most environments. Although, the library is fairly large (>100Mb), but is available on a CDN so need not stress server storage space. Most of this space comes from the bit-mapped fonts which are only downloaded on-demand, so should not stress web clients either. It also obviates the need for a TEX installation which `wp-latex` may require (although this plugin can use an external server also).

At face value, `mathjax-latex` necessarily adds very little semantics to the maths embedded within documents. The maths could be represented as `$$E=mc^2$$`, `\(E=mc^2\)]` or

```
<math> <mrow> <mi>E</mi> <mo>=</mo> <mrow> <mi>m</mi>
<msup> <mi>c</mi><mn>2</mn> </msup>
</mrow> </mrow> </math>
```

So, we have a heterogenous representation for identical knowledge. However, in practice, the situation is much better than this. The author of the work created these equation and has then read them, transformed by MathJax into a rendered form. If MathJax has failed to translate them correctly, in line with the author's intention, or if it has had some implications for the text in addition to setting the intended equations (if the TEX style markup appears accidentally elsewhere in the document), the author is likely to have seen this and fixed the problem. Someone wishing, for example, to extract all the mathematics as MathML from these documents computationally, therefore, knows:

- that the document contains maths as it imports MathJax
- that MathJax is capable of identifying this maths correctly
- that equations can be transformed to MathML using MathJax[7].

So, while our publication environment does not result directly in lower level of semantic heterogeneity, it does provide the data and the tools to enable the computational agent to make this transformation. While this is imperfect, it should help somewhat.

In short, we provide a practical mechanism to identify text containing mathematics and a mechanism to transform this to a single, standardised representation.

[7] This is assuming MathJax works correctly in general. The authors and readers are checking the rendered representation. It is possible that an equation would render correctly on screen, but be rendered to MathML inaccurately

4 Representing References

Unlike mathematics, there is no standard mechanism for reference and in-text citation, but there are a large number of tools for authors such as BibTeX, Mendeley [13] or EndNote. As a result of this, the integration with existing toolsets is of primary importance, while the representation of the in-text citations is not, as it should be handled by the tool layer anyway.

Within kblog, we have developed a plugin called kcite.[8] For the author, citations are inserted using the syntax:
`[cite]10.1371/journal.pone.0012258[/cite].`
The identifier used here is a DOI, or digital object identifier and, is widely used within the publishing and library industry. Currently, kcite supports DOIs minted by either CrossRef[9] or DataCite[10] (in practice, this means that we support the majority of DOIs). We also support identifiers from PubMed[11] which covers most biomedical publications and arXiv,[12] the physics (and other domains!) preprints archive, and we now have a system to support arbitrary URLs. Currently, authors are required to select the identifier where it is not a DOI.

We have picked this "shortcode" format for similar reasons as described for maths; it is relatively unambiguous, it is not XML based, so passes through the HTML generation layer of most authoring tools unchanged and is explicitly supported in WordPress, bypassing the need for regular expressions and later parsing. It would, however, be a little unwieldy from the perspective of the author. In practice, however, it is relatively easy to integrate this with many reference managers. For example, tools such as Zotero [14] and Mendeley use the Citation Style Language, and so can output kcite compliant citations with the following slightly elided code:

```
<citation>
   <layout prefix="[cite]" suffix="[/cite]"
        delimiter="[/cite] [cite]">
     <text variable="DOI"/>
   </layout>
 </citation>
```

We do not yet support LaTeX/BibTeX citations, although we see no reason why a similar style file should not be supported. We do, however, support BibTeX-formatted files: the first author's preferred editing/citation environment is based around these with Emacs, RefTeX, and asciidoc. While this is undoubtedly a rather niche authoring environment, the (slightly elided) code for supporting this demonstrates the relative ease with which tool chains can be induced to support kcite:

[8] http://wordpress.org/extend/plugins/kcite/
[9] http://wordpress.org/extend/plugins/kcite/
[10] http://www.datacite.org/
[11] http://www.ncbi.nlm.nih.gov/pubmed/
[12] http://arxiv.org/

```
(defadvice reftex-format-citation (around phil-asciidoc-around activate)
  (if phil-reftex-citation-override
      (setq ad-return-value (phil-reftex-format-citation entry format))
    ad-do-it))

(defun phil-reftex-format-citation( entry format )
  (let ((doi (reftex-get-bib-field "doi" entry)))
    (format "pass:[[cite source='doi'\\]%s[/cite\\]]" doi)))
```

The key decision with kcite from the authorial perspective is to ignore the reference list itself and focus only on in-text citations, using public identifiers to references. This simplifies the tool integration process enormously, as this is the only data that needs to pass from the author's bibliographic database onward. The key advantage for authors here is two-fold: they are not required to populate their reference metadata for themselves, and this metadata will update if it changes. Secondly, the identifiers are checked; if they are wrong, the authors will see this straightforwardly as the entire reference will be wrong. Adding DOIs or other identifiers moves from becoming a burden for the author to becoming a specific advantage.

While supporting multiple forms of reference identifier (CrossRef DOI, DataCite DOI, arXiv and PubMed ID) provides a clear advantage to the author, it comes at considerable cost. While it is possible to get metadata about papers from all of these sources, there is little commonality between them. Moreover, resolving this metadata requires one outgoing HTTP request[13] per reference, which browser security might or might not allow.

So, while the presentation of mathematics is performed largely on the client, for reference lists the kcite plugin performs metadata resolution and data integration on the server. A caching functionality is provided, storing this metadata in the WordPress database. The bibliographic metadata is finally transferred to the client encoded as JSON, using asynchronous call-backs to the server.

Finally, this JSON is rendered using the citeproc-js library on the client. In our experience, this performs well, adding to the readers' experience; in-text citations are initially shown as hyperlinks; rendering is rapid, even on aging hardware, and finally in-text citations are linked both to the bibliography and directly through to the external source. Currently, the format of the reference list is fixed, however, citeproc-js is a generalised reference processor, driven using CSL[14]. This makes it straight-forward to change citation format, at the option of the reader, rather than the author or publisher. Both the in-text citation and bibliography support outgoing links direct to the underlying resources[15]. As these links have been used to gather metadata, they are likely to be correct. While these advantages are relatively small currently, we believe that the use of JavaScript rendering over a linked references can be used to add further reader value in future.

[13] In practice, it is often more; DOI requests, for instance, use `303` redirects.

[14] http://citationstyles.org/

[15] Where the identifier allows – PubMed IDs redirect to PubMed.

For the computational agent wishing to consume bibliographic information, we have added significant value compared to the pre-formatted HTML reference list. First, all the information required to render the citation is present in the in-text citation next to the text that the authors intended. A computational agent can, therefore, ignore the bibliography list itself entirely. These primary identifiers are, again, likely to be correct because the authors now need them to be correct for their own benefit.

Should the computational agent wish, the (denormalised) bibliographic data used to render the bibliography is actually available, present in the underlying HTML as a JSON string. This is represented in a homogeneous format, although, of course, represents our (kcite's) intepretation of the primary data.

A final, and subtle, advantage of kcite is that the authors can only use public metadata, and not their own. If they use the correct primary identifier, and still get an incorrect reference, it follows that the public metadata must be incorrect[16]. Authors and readers therefore must ask the metadata providers to fix their metadata to the benefit of all. This form of data linking, therefore, can even help those who are not using it.

4.1 Microarray Data

Many publications require that papers discussing microarray experiments lodge their data in a publically available resource such as ArrayExpress [15]. Authors do this placing an ArrayExpress identifier which has the form `E-MEXP-1551`. Currently, adding this identifier to a publication, as with adding the raw data to the repository is no direct advantage to the author, other than fulfilment of the publication requirement. Similarly, there is no existing support within most authoring environments for adding this form of reference.

For the knowledgeblog-arrayexpress plugin,[17] therefore, we have again used a shortcode representation, but allowed the author to automatically fill metadata, direct from ArrayExpress. So a tag such as:
`[aexp id="E-MEXP-1551"]species[/aexp]`
will be replaced with *Saccharomyces cerevisiae*, while:
`[aexp id="E-MEXP-1551"]releasedate[/aexp]`
will be replaced by "2010-02-24". While the advantage here is small, it is significant. Hyperlinks to ArrayExpress are automatic, authors no longer need to look up detailed metadata. For metadata which authors are likely to know anyway (such as Species), the automatic lookup operates as a check that their ArrayExpress ID is correct. As with references 6, the use of an identifier becomes an advantage rather than a burden to the authors.

Currently, for the reader there is less significant advantage at the moment. While there is some value to the author of the added correctness stemming from the ArrayExpress identifier. However, knowledgeblog-arrayexpress is currently under-developed, and the added semantics that is now present could be used

[16] Or, we acknowledge, that kcite is broken!
[17] http://knowledgeblog.org/knowledgeblog-arrayexpress

more extensively. The unambiguous knowledge that:
`[aexp id="E-MEXP-1551"]species[/aexp]`
represents a species would allow us, for example, to link to the NCBI taxonomy database.[18]

Likewise, advantage for the computational agent from knowledgeblog-arrayexpress is currently limited; the identifiers are clearly marked up, and as the authors now care about them, they are likely to be correct. Again, however, knowledgeblog-arrayexpress is currently under developed for the computational agent. The knowledge that is extracted from ArrayExpress could be presented within the HTML generated by knowledgeblog-arrayexpress, whether or not it is displayed to the reader for, essentially no cost. By having an underlying short-code representation, if we choose to add this functionality to knowledgeblog-arrayexpress, any posts written using it would automatically update their HTML. For the text-mining bioinformatician, even the ability to unambiguously determine that a paper described or used a data set relating to a specific species using standardised nomenclature[19] would be a considerable boon.

5 Discussion

Our approach to semantic enrichment of articles is a measured and evolutionary approach. We are investigating how we can increase the amount of knowledge in academic articles presented in a computationally accessible form. However, we are doing so in an environment which does not require all the different aspects of authoring and publishing to be over-turned. More over, we have followed a strong principle of semantic enhancement which offers advantages to both reader and author immediately. So, adding references as a DOI, or other identifier, 'automagically' produces an in text citation and a nicely formatted reference list: that the reference list is no longer present in the article, but is a visualisation over linked data; that the article itself has become a first class citizen of this linked data environment is a happy by-product.

This approach, however, also has disadvantages. There are a number of semantic enhancements which we could make straight-forwardly to the knowledgeblog environment that we have not; the principles that we have adopted requires significant compromise. We offer here two examples.

First, there has been significant work by others on CiTO [16] – an ontology which helps to describe the relationship between the citations and a paper. Kcite lays the ground-work for an easy and straight-forward addition of CiTO tags surrounding each in-text citation. Doing so, would enable increased machine understandability of a reference list. Potentially, we could use this to the advantage to the reader also: we could distinguish between reviews and primary research papers; highlight the authors' previous work; emphasise older papers which are being refuted. However, to do this requires additional semantics from the author. Although these CiTO semantic enhancements would be easy to insert

[18] http://www.ncbi.nlm.nih.gov/Taxonomy/
[19] the standard nomenclature was only invented in 1753 and is still not used universally.

directly using the shortcode syntax, most authors will want to use their existing reference manager which will not support this form of semantics; even if it does, the author themselves gain little advantage from adding these semantics. There are advantages for the reader, but in this case not for both author and reader. As a result, we will probably add such support to kcite; but, if we are honest, find it unlikely that when acting as content authors, we will find the time to add this additional semantics.

Second, our presentation of mathematics could be modified to automatically generate MathML from any included TEX markup. The transformation could be performed on the server, using MathJax; MathML would still be rendered on the client to webfonts. This would mean that any embedded maths would be discoverable because of the existence of MathML, which is a considerable advantage. However, neither the reader nor the author gain any advantage from doing this, while paying the cost of the slower load times and higher server load that would result from running JavaScript on the server. More over, they would pay this cost regardless of whether their content were actually being consumed computationally. As the situation now stands, the computational user needs to identify the insert of MathJax into the web page, and then transform the page using this library, none of which is standard. This is clearly a serious compromise, but we feel a necessary one.

Our support for microarrays offers the possibility of the most specific and increased level of semantics of all of our plugins. Knowledge about a species or a microarray experimental design can be very prescisely represented. However, almost by definition, this form of knowledge is fairly niche and only likely to be of relevance to a small community. However, we do note that the knowledgeblog process based around commodity technology does offer a publishing process that can be adapted, extended and specialised in this way relatively easily. Ultimately the many small communities that make up the long-tail of scientific publishing adds up to one large one.

6 Conclusion

Semantic publishing is a desirable goal, but goals need to be realistic and achievable. to move towards semantic publishing in kblog, we have tried to put in place an approach that gives benefit to readers, authors and computational interpretation. As a result, at this stage, we have light semantic publishing, but with small, but definite benefits for all.

Semantics give meaning to entities. In kblog, we have sought benefit by "saying" within the kblog environment that entity x is either **maths**, a **citation** or a **microarray** data entity reference. This is sufficient for the kblog infra-structure to "know what to do" with the entity in question. Knowing that some publishable entity is a "lump" of maths tells the infra-structure how to handle that entity: the reader has benefit from it looking like maths; the author has benefit by not having to do very much; and the infra-structure knows what to do. In addition, this approach leaves in hooks for doing more later.

It is not necessarily easy to find compelling examples that give advantages for all steps. Adding in CiTO attributes to citations, for instance, has obvious advantages for the reader, but not the author. However, advantages may be indirect; richer reader semantics may give more readers and thus more citations—the thing authors appreciate as much as the act of publishing itself. It is, however, difficult to imagine how such advantages can be conveyed to the author at the point of writing. It is easy to see the advantages of semantic publishing for readers, as a community we need to pay attention to advantages to the authors. Without these "carrots", we will only have "sticks" and authors, particularly technically skilled ones, are highly adept at working around sticks.

References

1. Shadbolt, N., Hall, W., Berners-Lee, T.: The semantic web revisited. Intelligent Systems, IEEE **21**(3) (2006) 96–101 1
2. Shotton, D.: Semantic publishing: the coming revolution in scientific journal publishing. Learned Publishing **22**(2) (2009) 85–94 1
3. Shotton, D., Portwin, K., Klyne, G., Miles, A.: Adventures in semantic publishing: exemplar semantic enhancements of a research article. PLoS computational biology **5**(4) (2009) e1000361 1
4. Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. International Journal on Semantic Web and Information Systems (IJSWIS) **5**(3) (2009) 1–22 1
5. Landport, L.: The LaTeX book. Adison wesley, Reading, MA (1984) 1
6. Lord, P., Cockell, S., Swan, D.C., Stevens, R.: The ontogenesis knowledgeblog: Lightweight publishing about semantics, with lightweight semantic publishing. In: Semantic Web Technologies for Libraries and Readers. (2011) 2, 3
7. Wordpress: http://www.wordpress.org. 2
8. Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M.R., Li, P., Oinn, T.: Taverna: a tool for building and running workflows of services. Nucleic Acids Res **34**(Web Server issue) (Jul 2006) 729–732 2
9. Knuth, D.E.: The TeX Book. 3rd edition edn. Adison Wesley, Reading, MA (1986) 4
10. WP Latex: http://wordpress.org/extend/plugins/wp-latex/. 4
11. Mathjax-Latex: http://wordpress.org/extend/plugins/mathjax-latex/. 4
12. Mathjax: http://www.mathjax.org. 6
13. Mendeley: http://www.mendeley.org. 7
14. Zotero: http://www.zotero.org. 7
15. Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G.G., Oezcimen, A., Rocca-Serra, P., Sansone, S.A.: ArrayExpress- public repository for microarray gene expression data at the EBI. Nucleic Acids Research **31**(1) (2003) 68–71 9
16. Shotton, D.: CiTO, the Citation Typing Ontology. Journal of Biomedical Semantics **1**(Suppl 1) (2010) S6 10

Online open neuroimaging mass meta-analysis

Finn Årup Nielsen[1]*, Matthew J. Kempton[2], and Steven C. R. Williams[2]

[1] DTU Informatics, Technical University of Denmark, Lyngby, Denmark.
fn@imm.dtu.dk, http://www.imm.dtu.dk/~fn/
[2] Department of Neuroimaging, Institute of Psychiatry, King's College London, London, UK

Abstract. We describe a system for meta-analysis where a wiki stores numerical data in a simple format and a web service performs the numerical computation. We initially apply the system on multiple meta-analyses of structural neuroimaging data results. The described system allows for mass meta-analysis, e.g., meta-analysis across multiple brain regions and multiple mental disorders.

1 Introduction

The scientific process aggregates a large number of scientific results into a common scientific consensus. *Meta-analysis* performs the aggregation by statistical analysis of numerical values presented across scientific papers. Collaborative systems such as wikis may easily aggregate text and values from multiple sources. However, so far they have had limited ability to apply numerical analysis as required, e.g., by meta-analysis.

Researchers have discussed the advantages and disadvantage of the tools for conducting systematic reviews from "paper and pencil", over spreadsheets to RevMan and web-based specialized applications [10]: Setup cost, versatility, ability to manage data, etc. In 2009 they concluded that "no single data-extraction method is best for all systematic reviews in all circumstances". For example, RevMan and Archie of the Cochrane Library provide an elaborate system for keeping track and analyzing textual and numerical data in meta-analyses, but the system could not import information from electronic databases [10]. Our original meta-analyses [4, 5] relied on the Microsoft Excel spreadsheets later distributed on public web sites. Compared to an ordinary spreadsheet a wiki solution provides data entry provenance and collaborative data entry with immediately update. Shareable folders on cloud-based storage systems would help collaboration on spreadsheets, but yield no provenance. Online services, such as the spreadsheet of Google Docs, may lack meta-analytic plotting facility. Web-based specialized applications for systematic reviews may have a high setup cost [10].

We have previously explored a simple online meta-analysis system—a "fielded wiki"—in connection with personality genetics [8]. As implemented specifically

* Thanks to the Lundbeck Foundation for the funding of the *Center for Integrated Molecular Brain Imaging* (CIMBI).

for this scientific area the web service lacks generality for other types for meta-analytic data. Furthermore the system relied on PubMed or Brede Wiki to represent bibliographic information.

Following Ward Cunningham's quote "What's the simplest thing that could possibly work?" we present a simple system that allows for mass meta-analysis of numerical data presented as comma-separated values (CSV) in a standard MediaWiki-based wiki, — the Brede Wiki: `http://neuro.imm.dtu.dk/wiki/`.

2 Data and data representation

We use the MediaWiki-based Brede Wiki to represent the data [1]. For our neuroimaging data each data record usually consists of three values (number of subjects, their mean and standard deviation). The individual study typically compares two such data records, e.g., from a patient and a control group. We also record labels for the data record, e.g., the biographic information, as well as extra subject information about the two groups, such as age, gender and clinical characteristics, so that the total number of data items for each study may be seven or more. Each meta-analysis will usually determine what extra relevant information should be included and it may differ between studies, e.g., a *Y-BOCS* value has typically only relevance for obsessive-compulsive disorder patients. The functional neuroimaging area has *CogPO* and *Cognitive Atlas* ontologies enabling researchers to describe the topic of an experiment, but these efforts do not directly apply to our data. One CSV line carries the information for each study.

Separate wiki pages store— rather than uploaded files— the CSV data, so the MediaWiki template functionality can transclude the CSV data on other wiki pages. By convention pages with CSV information have the ".csv" extension as part of the title so external scripts can recognize them as special pages and the wiki pages have no wiki markup.

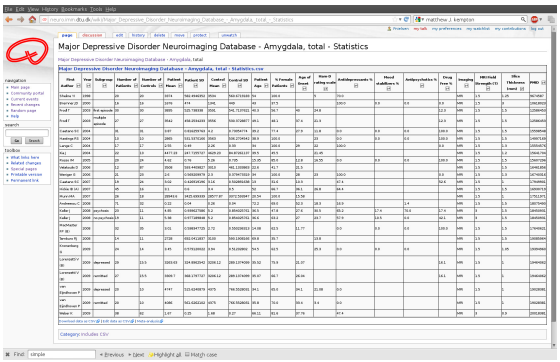


Fig. 1. Screenshot from the wiki showing CSV data transcluded on a page.

MediaWiki templates may generate links for download, editing and meta-analysis of the data. Presently, no controlled vocabulary beyond the template fields describes the columns in the CSV. To generate an appropriate content-type (text/csv) a bridging web script functions as a proxy, so a download of the CSV page can spawn a client-side spreadsheet program.

A few MediaWiki extensions can format CSV information: *SimpleTable* and *TableData*. Figure 1 shows the transclusion of CSV data with a modified version of the SimpleTable extension. The Brede Wiki uses the standard template

system for recording structured bibliographic data about the publication and to annotate the CSV information, see Figure 2.

```
{{Metaanalysis csv begin}}
{{Metaanalysis csv
 | title = Major Depressive Disorder Neuroimaging Database - Pituitary, total
 | topic1 = Pituitary
 | topic2 = Major depressive disorder
 | topic3 = MaND
}}
{{Metaanalysis csv
 | title = Obsessive-compulsive disorder Neuroimaging Database - Pituitary
 | topic1 = Pituitary
 | topic2 = Obsessive-compulsive disorder
 | topic3 = ObND
}}
{{Metaanalysis csv end}}
```

Fig. 2. Template to annotate the CSV data and define the links to the meta-analysis.

The bulk of the data currently presented in the wiki comes from the large mass meta-analysis of volumetric studies on major depressive disorder reporting over 50 separate meta-analyses for individual brain regions [4]. Further data comes from mass meta-analyses across multiple brain regions on bipolar disorder [5] and first-episode schizophrenia [6], a meta-analysis on longitudinal development in schizophrenia [7] as well as data from individual original studies on obsessive-compulsive disorder.

Apart from neuroimaging studies the Brede Wiki also records data from meta-analyses from a few other studies outside neuroimaging [2], allowing us to test the generality of the framework. The data is distributed under ODbL.

3 Web script and meta-analysis

The web script for meta-analysis reads the CSV information, identifies the required columns for meta-analysis, performs the statistical computations and makes meta-analytic plots— the so-called forest and funnel plots—in the SVG format, see Figure 3. From either the title information or a PubMed identifier the script generates back-links from the generated page to pages on the wiki. The script may also export the computed results as JSON or CSV. Furthermore, it may generate a small *R* script that sets up the data in variables and use the `meta` library for meta-analysis.

The web script attempts to guess the separator used on the CSV page and also tries to match the elements of the column header, e.g., the strings "control n", "controls number", "number of controls", etc. match for number of control subjects. With no matches the user needs to explicitly specify the relevant columns via URL parameters, which in turn a wiki template can setup.

Standard meta-analysis computes an *effect size* from each result in a paper and computes a combined meta-analytic effect size and its confidence interval. Although the methodological development continues, there exist established statistical analysis approaches for ordinary meta-analysis [2]. Our system implements computations on the standardized mean difference for continuous variables and on the natural logarithm of the odds ratio for categorical variables with fixed and random effects methods using an inverse-weighted variance model, — following the approach in the Stata program. As an extra option we provide meta-analysis on the natural logarithm of the variance ratio [3], for comparison of the standard deviations between two groups of subjects.

4 Results

We have added 124 pages with CSV data, — most of which contain data suitable for meta-analysis. For individual analyses the reading, computation and download finish within seconds. With multiple calls to the web script and JSON output another script can plot multiple meta-analytic results together as in Figure 4. Generating such a plot takes several minutes. For generating the page show in Figure 3 we need only the CSV data and the web script, while the script that generated Figure 4 used information defined in templates, CSV data and the web script with no further adaption of MediaWiki.

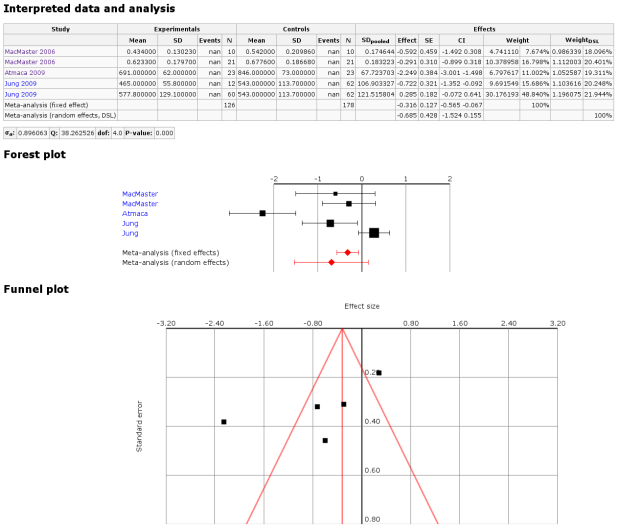


Fig. 3. Screenshot of web script showing the meta-analytic results with forest and funnel plots.

5 Discussion

By using MediaWiki in our present system we exploit the template facility to capture structured information, and free-form wikitext for annotation and comment on the individual scientific papers, — as in semantic academic annotation wikis *AcaWiki* and *WikiPapers*. It is also possible to use the pages of the wiki as a simple means to keep track of the status of the papers considered for the meta-analysis: potentially eligible, eligible, partially entered and fully entered.

Why not Semantic MediaWiki? Semantic MediaWiki (SMW) may query text and numerical data, though has not had the ability to make complex computations. The *Semantic Result Formats* extension includes average, sum, product and count result formats enabling simple computations of a series of numerical values, but insufficient for the kind of computations we require. The data for meta-analysis form a n-ary data record (mean, standard deviation, number of subjects, labels) so either individual SMW pages should store each data record or we should invoke the n-ary functionality in *Semantic Internal Objects* SMW extension, SMW *record* or the recently-introduced *subobject* SMW functionality. We have not investigated whether these tools provide convenient means for representing our data. The Brede Wiki can export its ontologies defined in MediaWiki template to SKOS. Our future research can consider RDFication of the CSV information through the SCOVO format [9].

We wrote the web service in Python, where Numpy makes vector computation available and Scipy provides statistical methods, necessary for the computation. In a future PHP implementation the script could more closely integrate with the wiki as either a MediaWiki or a SMW extension.

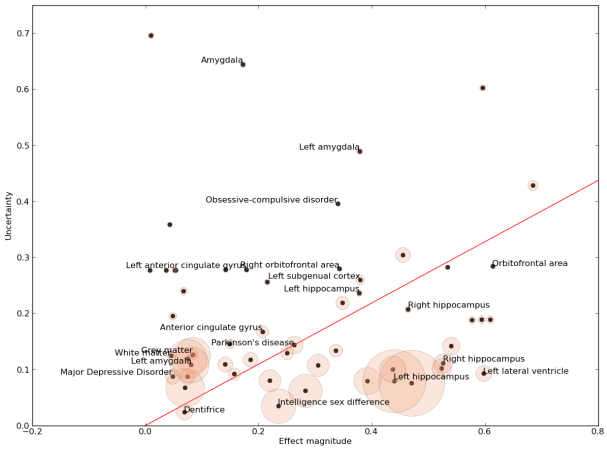


Fig. 4. Results from mass meta-analyses shown in a L'Abbé-like plot and constructed by calling the web script multiple times. Each dot corresponds to a meta-analysis. Uncertainty as a function of effect size with size of each dot determined by the number of subjects. The line indicates 0.05-significance.

A wiki built from standard components provides a inexpensive solution with means to manage meta-analytic data in a collaborative environment. The general framework allows not only the meta-analysis of neuroimaging-derived data but has the potential for managing and analyzing data from many other domains.

References

1. Nielsen, F.Å.: Brede Wiki: Neuroscience data structured in a wiki. SemWiki (2009) 129–133
2. Hartung, J. et al.: Statistical Meta-Analysis with Applications. (2008)
3. Shaffer, J.P.: Caution on the use of variance ratios: a comment. Review of Educational Research **62**(4) (1992) 429–432
4. Kempton, M.J. et al.: Structural neuroimaging studies in major depressive disorder: meta-analysis and comparison with bipolar disorder. Archives of General Psychiatry **68**(7) (2011) 675–690
5. Kempton, M.J. et al.: Meta-analysis, database, and meta-regression of 98 structural imaging studies in bipolar disorder. Arch. Gen. Psychiatry **65**(9) (2008) 1017–1032
6. Steen, R.G. et al.: Brain volume in first-episode schizophrenia: systematic review and meta-analysis of magnetic resonance imaging studies. Brit. J. Psychiatry **188** (2006) 510–518
7. Kempton, M.J. et al.: Progressive lateral ventricular enlargement in schizophrenia: a meta-analysis of longitudinal MRI studies. Schiz. Res. **120**(1–3) (2010) 54–52
8. Nielsen, F.Å.: A fielded wiki for personality genetics. WikiSym (2010)
9. Hausenblas, M. et al.: SCOVO: Using statistics on the web of data. ESWC (2009) 708–722
10. Elamin, M.B. et al.: Choice of data extraction tools for systematic reviews depends on resources and review complexity. J. Clin. Epidemiology **62**(5) (2009) 506–510

Uncovering impacts: a case study in using altmetrics tools

Jason Priem[1], Cristhian Parra[2], Heather Piwowar[3], Paul Groth[4], and Andra Waagmeester[5]

[1] University of North Carolina at Chapel Hill
priem@email.unc.edu
[2] University of Trento
parra@disi.unitn.it
[3] National Evolutionary Synthesis Center
hpiwowar@nescent.org
[4] VU University Amsterdam
p.t.groth@vu.nl
[5] Maastricht University
andra.waagmeester@maastrichtuniversity.nl

Abstract. Growing scholarly use of Web tools present an opportunity to track alternative impacts along heretofore invisible paths like reading, bookmarking, and discussing. We present two tools, CitedIn and total-impact, that gather and report these and other "altmetrics" After discussing the tools features, we use a set of 214 articles from a national research center as a demonstration case study. We find that both tools present a meaningful number and variety of altmetrics in a form that could be used for immediate evaluation, and call for more research into the properties and validity of altmetrics.

Keywords: altmetrics, scholarly communication, impact, tools

1 Introduction

The future of scholarly communication is one in which a large part of scholarly communication is conducted online [3]. A key part of the scholarly communication lifecycle is trying to understand the impact of work. The process of understanding impact helps scientists, science administrators and others both find, evaluate, and access scholarly products. Traditionally, this impact assessment has been done primarily through the tracking of formal citations. This is possible because citations counts, for all their occasional ambiguity [2], do reflect use of scholarly products. However, this reflection is of a restricted spectrum; scholarly products are often used by scholars, and others, in ways that do not perturb the citation record [5]. Furthermore, traditional citation does not reflect the rapid nature of communications afforded by the Web. Thus, we need new approaches for measuring impact in this changed world.

Indeed, because of the Web scholarly communication, formerly "underground" uses like reading, bookmarking, sharing, discussing, and rating are beginning to leave online traces. The are becoming visible on Web pages [8, 13],

on blogs [6], in downloads [1, 4], on social media like Twitter [9], and in social reference managers like CiteULike, Mendeley, and Zotero [7]. These alternatives to traditional citation analysis have been labeled altmetrics [11]. Altmetrics offer potential for gathering information on more diverse types of impact, from more diverse scholarly products, including blog posts, slides, datasets, or even tweets. They also have the important benefit of speed; altmetrics typically accumulate in days or weeks rather than the years citations require. This is particular useful in as the research process increases pace where users of scientific content need to understand the impact of it rapidly. To begin to make practical use of altmetrics for measuring impact requires both a greater understanding of the properties and validity of these new metrics, and practical tools for obtaining them [10]. Others have begun the former [12]; here we will pursue the latter, presenting two new tools for gathering and presenting altmetrics.

2 Tools for Altmetrics: CitedIn and total-impact

CitedIn (http://citedin.org) and total-impact (http://total-impact.org) are open-source tools that receive as input a list of identifiers for scholarly products, and output a set of altmetrics for each product. CitedIn accepts only articles with PubMed IDs (PMIDs); total-impact accepts articles identified by PMID or DOI, but also datasets and slides using a variety of identifiers including URL, handle, and accession numbers. Both tools allow users to input identifiers manually; CitedIn also offers a REST API, and total-impact lets users automatically populate the products list using items stored in Mendeley or Slideshare libraries. Once users have uploaded products, CitedIn and total-impact both use calls to open Web APIs to gather data about them; CitedIn also caches available databases. As of September 25, 2011, the data sources used by each are listed in Table 1.

In addition to gathering altmetrics from these sources, both tools also include some additional features. CitedIn lets users input and output data over a REST API, and also reports a "CI-number" that summarizes all almetrics activity in a single value. Total-impact offers persistent URLs for impact report pages; the impact metrics can be refreshed over time. Both tools let users download results as structured text files for further analysis. Output pages for the tools are shown in Figures 1 and 2.

3 Case study: altmetrics for a national research center

We used a set of 214 articles from the National Evolutionary Synthesis Center (NESCent) as a realistic test for the two tools. NESCent was interested in tracking the impact of work they funded in a faster and more comprehensive way than citation analysis allowed – a typical use case for altmetrics. We entered the articles into CitedIn on August 14 2011, and into total-impact September 23 2011, then collected and analyzed the results.

All 214 articles had DOIs, and so were able to be processed by total-impact. Only 174 articles had the PMIDs required by CitedIn, so the CitedIn sample

	CitedIn	total-impact
Data repositories, including locating datasets associated with a given publication	ABS, Ares, Alzgene, Biogrid, BredeWiki, Ctdatabase, cancerCell, ChdWiki, Cosmic, Ctd, Cutdb, Dejavu, HIFTFBS, HNF4, HaemB, Jaspar, Kegg, Mgi, Mint, Mpidb, Nfi Regulome Resource, Oreganno, MIDNCI, BIDReactome, BDB, PleiadesGenes, Gregransbase, Balmer Retinoic, Uniprot, Wikipathways, Wormbase, YTPdb, Zfin	Dryad (downloads of most popular file, package views, total downloads and file views)
Social bookmarking and reference management tools	CiteULike, Connotea, Mendeley	CiteULike, Delicious, Mendeley (groups, readers)
Blogs and social media	Google Blogs, Nature Blogs	Facebook (clicks, comments, likes, shares)
Traditional citation	Google Books mentions	Citation in PubMed Central
Other	PubMed subsets, Citations from Wikipedia (pmid)	Citations from Wikipedia
PLoS ALM	N/A	Connotea, citations (CrossRef, PubMed Central, Scopus), blog mentions(Nature Blogs, ResearchBlogging, Bloglines, Postgenomic), downloads, PubMed activity

Table 1. Data sources for CitedIn and total-impact as of September 2011

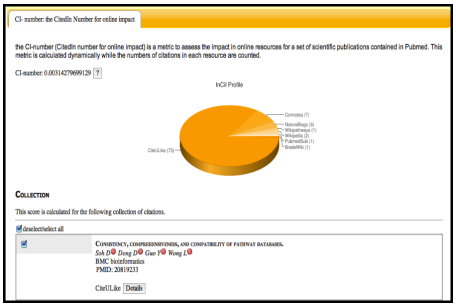


Fig. 1. CitedIn results page (Sept 2011)

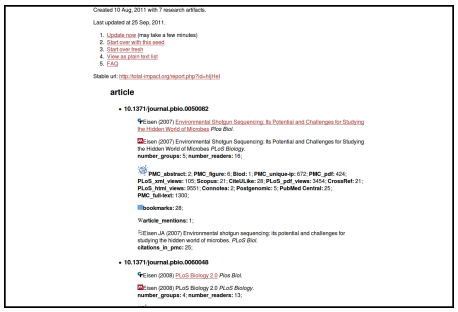


Fig. 2. total-impact results (Sept 2011)

is smaller. Both tools showed that altmetric activity as measured by number of "altmetric events" (bookmarks, downloads, etc.) is relatively widespread across articles: CitedIn found at least one event on 95% of its articles, and total-impact on 85%. There were a mean of 28 and median of 16 events per CitedIn article, with a maximum of 678. Total-impact had a per-article mean of 92 events and a median of 19; the higher mean is due to Dryad dataset downloads, which accumulate more easily than other metrics, reaching a maximum of 2769 on one article. We visualized the activity across articles using heatmaps, shown in Figures 3 and 4 to create a sort of "impact genome." Only altmetrics with nonzero counts are shown, and counts of each altmetric are normalized by that metric's maximum. Articles are arranged so that those with higher mean event counts across all metrics are further left.

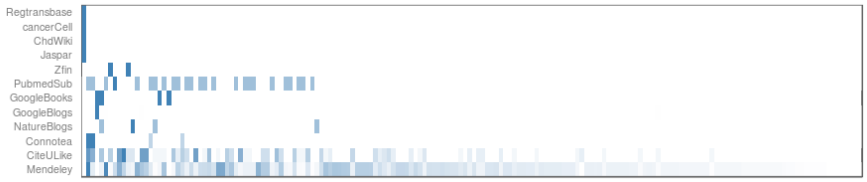


Fig. 3. Active CitedIn event types and normalized event counts per article.

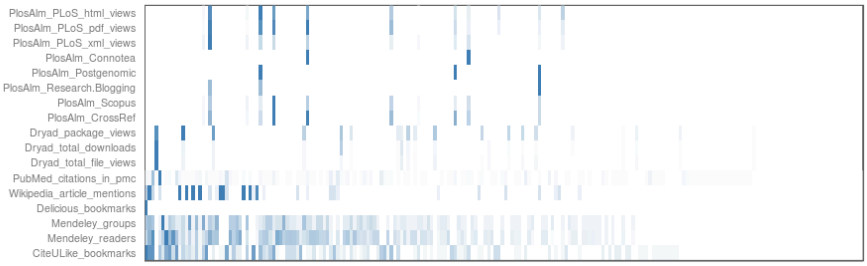


Fig. 4. Active total-impact event types and normalized event counts per article.

4 Conclusion

Altmetrics have potential to improve the speed and breadth of scientific evaluation. CitedIn and total-impact are two tools in early development that aim to gather altmetrics. A test of these tools using a real-life dataset shows that they work, and that there is a meaningful amount of altmetrics data available for use. These tools continue to improve: check out the current versions for up to date capabilities.

The properties and validity of these data, however, are still unclear, and call for additional research. What is the scholarly value of, for instance, a Mendeley

bookmark or a Wikipedia citation? Future work should also investigate how altmetrics for different sets of articles can be compared; this is a particularly tricky problem given the high dimensionality of altmetrics data, and may benefit from better visualization techniques, or statistical approaches like principle component analysis and factor analysis.

- Source code for CitedIn: http://code.google.com/p/citedin
- Source code for total-impact: https://github.com/mhahnel/total-impact
- Source code and data for analysis in this paper:
  https://github.com/jasonpriem/altmetrics-tools-iConference-poster
- The authors of the paper are key developers on CitedIn and Total-Impact

References

1. Bollen, J., Van de Sompel, H., Hagberg, A., Chute, R.: A principal component analysis of 39 scientific impact measures. PLoS ONE 4(6), e6022 (06 2009)
2. Bornmann, L., Daniel, H.D.: What do citation counts measure? A review of studies on citing behavior. Journal of Documentation 64(1), 45–80 (2008)
3. Bourne, P., Clark, T., Dale, R., de Waard, A., Herman, I., Hovy, E., Shotton, D., on behalf of the Force11 community: Force11 White Paper: Improving the Future of Research Communication and e-Scholarship (2011), `http://force11.org/white_paper`
4. Brody, T., Harnad, S., Carr, L.: Earlier web usage statistics as predictors of later citation impact: Research articles. J. Am. Soc. Inf. Sci. Technol. 57(8), 1060–1072 (Jun 2006), `http://dx.doi.org/10.1002/asi.v57:8`
5. Cronin, B.: The Hand of Science: Academic Writing and Its Rewards. Scarecrow Press (2005)
6. Groth, P., Gurney, T.: Studying Scientific Discourse on the Web using Bibliometrics: A Chemistry Blogging Case Study. In: WebSci10 Extending the Frontiers of Society OnLine (2010)
7. Hull, D., Pettifer, S.R., Kell, D.B.: Defrosting the digital library: bibliographic tools for the next generation web. PLoS computational biology 4(10), e1000204 (2008), `http://www.ncbi.nlm.nih.gov/pubmed/18974831`
8. L., V., K., H.: Relationship between links to journal web sites and impact factors. Aslib Proceedings: new information perspectives 54(6), 356–361 (2002)
9. Priem, J., Costello, K.L.: How and why scholars cite on Twitter. Proceedings of the 73rd ASIS&T Annual Meeting 73, `http://jasonpriem.org/self-archived/Priem\_Costello\_Twitter.pdf`
10. Priem, J., Hemminger, B.H.: Scientometrics 2.0: New metrics of scholarly impact on the social Web. First Monday 15(7) (Jul 2010), `http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2874`
11. Priem, J., Taraborelli, D., Groth, P., Neylon, C.: Alt-metrics: a manifesto (2010), `http://altmetrics.org/manifesto/`
12. Shema, H., Bar-Ilan, J.: Characteristics of Researchblogging.org science Blogs and Bloggers. altmetrics 11, `http://altmetrics.org/workshop2011/shema-v0/`
13. Thelwall, M., Vaughan, L., Björneborn, L.: Webometrics. Annual Review of Information Science and Technology 39(1), 81–135 (Oct 2006), `http://dx.doi.org/10.1002/aris.1440390110`

Semantic Publishing of Knowledge about Amino Acids

Robert Stevens[1] and Phillip Lord[2]

[1] [1]School of Computer Science, University of Manchester, UK
[2] [2]Department of Computing Science, University of newcastle, Newcastle, UK
Robert.Stevens@Manchester.ac.uk

Abstract. We semantically publish knowledge about the amino acids commonly described within biochemistry. We do this as an ontology written in OWL and presented as XML/RDF. The classification of amino acids is based on taylor's article (PMID:3461222) from 1986 published in the Journal of Theoretical Biology. The ontology goes further than the static paper version; it combines many aspects of the physicochemical properties taylor uses to classify amino acids to give a rich, multi axial classification of amino acids. Taylor's original description of the amino acid's physicochemical properties are captured with value partitions and restrictions on the amino acid classes themselves. A series of defined classes then establishes the multi-axial classification. The publication, hwen loaded into an OWL ontology manipulation tool, allows some knowledge about amino acids to be explored and used computationally. By publishing this knowledge about amino acids as a semantic document in the form of an ontology we persue an agenda of disruptive technology in publishing. It allows us to 'push' at the nature of a semantic publication.

Blogs about the published semantics of amino acids may be found at http://robertdavidstevens.wordpress.com/2010/12/18/an-update-to-the-amino-acids-ontology/ and links following. The ontology is at http://www.cs.man.ac.uk/~stevensr/ontology/amino-acids.owl.

It is, perhaps, an ontologist's question to ask 'what is semantic publishing?". When is a publication semantic and when does some computational semantic artefact become a publication? A further question is when is a semantic publication a scientific publication? Our submission to Sepublica 2012 was an experiment in this area—or it was the authors 'just trying it on'. Whichever it is, the reviewers have gone along with our game, so here's a narrative around our submission of an ontology of amino acids written in the Web Ontology Language (OWL) as a semantic publication to sepublica 2012. This narrative is a side-effect of our attempt at semantic publishing—we used our ontology of amino acids as a semantic publication, but does that count?; what is actually published and what can actually be read? this short text is really the front-end to our amino acids semantic publication, but it turns out that the narrative it provides, or something like it, is a necessary part of (scientific) semantic publishing.

A snippet from the Sepublica 'instructions to authors'[3] shows the origins of our submission:

> We also invite submissions in XHTML+RDFa or in the format or YOUR semantic publishing tool. However, to ensure a fair review procedure, authors must additionally export them to PDF.

this made us ask 'what would happen if we submited an RDF document for one of our ontologies as a submission to Sepublica?'. Our reasoning went something like this:

- Sepublica can have exactly what they've asked for...
- An ontology in OWL has an RDF syntax, so it matches the representation criterion;
- it has a URI that means it is published on the web, so it matches the publication criterion;
- The ontology captures some knowledge about a field of interest—that is, the semantics of that field, so it matches the semantic criterion;
- The ontology can be argued to be a document...

so, that ontology is a semantic publication. Anyway, we decided to 'try it on' and, to their credit, both the workshop organisers (after a query to find out if we'd done what we meant to do) and the reviewers went along with what we did. Given that the Amino Acids Ontology, in its RDF form, was accepted as a publication for Sepublica, we can conclude that it is a semantic publication.

Another interesting aspect of the Sepublica process is that the instructions asked for a PDF submission (in addition to any semantic submission) to ease the reviewing process. So, partly because the EasyChair site for sepublica was only set upto submit PDF and to take the organisers at their word, we first submitted a Manchester OWL Syntax version of the ontology and converted it into PDF. MOS is a more or less human readable syntax for OWL. However, the PDF version of the MOS wasn't especially useful. So, I asked for Easychair to be set up to allow non-PDF submission; it turns out that a zip file was the only way of achieving submission of an RDF document. If we are going to have semantic scientific publications, then we need a way of handling them; not just in Easychair's reviewing process, but in the wider context of the scientific workflow.

The blogs above give sufficient background for the ontology, but here is an outline. The Amino Acid ontology is a simple ontology that captures some basic conceptualisations of amino acids used by biochemists [1]. It has the basic criteria by which biochemists classify amino acids—size, polarity, charge, aromaticity and hydrophobicity. Only the biologically used amino acids are allowed and there are various constraints on the qualities permited for the amino acids. It works both as an exemplar of the role of automated reasoning in ontology maintenance and as a 'guide' to the amino acids. A coplex hierarchy of the amino acids is

[3] `http://sepublica.mywikipaper.org/drupal/node/23` accessed March 14 2012.

then offered, including some types of amino acid that cannot exist.[4] Thus the ontology captures the semantics of amino acid entities in some computational form over which reasoning can be performed. The ontology can be browsed using some OWL enabled tool and it can act as an amino acid 'tutorial' as well as supply computational semantics about amino acids to applications. This form of publication of Taylor's classification offers more than the original paper in terms of explicitness, computational manipulation and flexibility. It does, however, lack some, to say the least, of the context and narrative needed for a scientific publication.

We offered our submission to Sepublica as a *disruptive technology*[5]; semantic publishing should be a disruptive technology by creating a new publication market and changing the values by which scientific publishing happens. That Sepublica stil needs to ask for PDF to enable review (though the reviewers of the Amino Acids Ontology managed without) means that publication has not been disrupted enough; data are available with some computational semantics, but we don't have semantic scientific publication.

What does all of this tell us? The Amino Acids Ontology is a semantic publication, but also that it isn't really and it isn't a scientific semantic publication. While the ontology captures the semantics of the domain in a computational form, it lacks the narrative that semantic publication of data needs to make it useful for humans. We don't want the reverse of the current situation of all narrative and no computation, to be replaced by all computational semantics and no human narative. As others have already said, semantic publication needs human narrative. This would make most of the RDF only link data publications of scientific data only partially a semantic publication; linked data is necessary but not sufficient for semantic publication.

References

1. Taylor, W.R.: The classification of amino acid conservation. Journal of Theoretical Biology **119**(2) (1986) 205–218

[4] Links to supplementary information and the ontology itself are available in the abstract.

[5] `http://en.wikipedia.org/wiki/Disruptive_innovation`

Linked Data for the Natural Sciences: Two Use Cases in Chemistry and Biology

Cord Wiljes and Philipp Cimiano

Semantic Computing, CITEC, Bielefeld University, Germany
{cwiljes,cimiano}@cit-ec.uni-bielefeld.de
http://sc.cit-ec.uni-bielefeld.de

Abstract. The Web was designed to improve the way people work together. The *Semantic Web* extends the Web with a layer of *Linked Data* that offers new paths for scientific publishing and co-operation. Experimental raw data, released as Linked Data, could be discovered automatically, fostering its reuse and validation by scientists in different contexts and across the boundaries of disciplines. However, the technological barrier for scientists who want to publish and share their research data as Linked Data remains rather high. We present two real-life use cases in the fields of chemistry and biology and outline a general methodology for transforming research data into Linked Data. A key element of our methodology is the role of a *scientific data curator*, who is proficient in Linked Data technologies and works in close co-operation with the scientist.

Keywords: Research Data Management, Scientific Publishing, E-Science, Semantic Web, Ontology, Linked Data, Methodology.

1 Motivation

The World Wide Web was envisioned by its inventor Tim Berners-Lee as a universal information space that enables people to work together and collaborate better [4]. The *Semantic Web* adds an additional layer of *Linked Data* to the Web that allows machines to process the semantics of the data. The Semantic Web has the potential to change the way scientists co-operate and communicate, how they share data, and how they publish their research results. Because science has become more interdisciplinary, the need for the exchange of data between different branches of science has increased dramatically. The Semantic Web offers a solution to this challenge. Data from different fields could be combined in new ways, giving new insights and helping to solve complex problems that require an interdisciplinary approach.

A cornerstone of the scientific method is the requirement that any experiment has to be reproducible [16]. Publishing research data in an open fashion would support for instance:

– the discovery of related datasets, allowing for comparison of results in different contexts, obtained under different experimental conditions etc. This

requires that the data is published in some standard format (e.g. RDF) so that Semantic Web search engines can index all data and retrieve and rank all available datasets relevant for a given scientific question or research hypothesis.
– the external validation of data and reproduction of results by other parties. This requires that the data is sufficiently annotated and documented so that the exact experimental conditions can be identified.

Because scientific research data is very valuable, funding agencies have a high interest to prevent duplication of effort and foster the reuse existing data as efficiently as possible [2]. Nowadays, however, primary research data is still mostly stored in closed, non-accessible silos, usually on local hard discs in the scientist's lab. Typically, only the interpreted and aggregated results are made available to the scientific community via standard publication channels (e.g. journal and conference papers).

In general, scientists have been rather reluctant to adopt Semantic Web technologies. The reasons for this reluctance were revealed by several surveys (for a summary cf. [11]). Presumably the most important one is the lack of incentives, i.e. there is so far only limited reward and recognition for publishing research data. In addition, scientists often regard the results of their research as their property and fear others might take unfair advantage of it. Especially in highly competitive research areas this is a major concern.

But there is a growing number of scientists who share the ideal of making research data public and are willing to publicly release their data. These early adopters face another barrier in the form of technical complexity. A considerable effort is necessary to get acquainted with the relevant techniques and paradigms, i.e. Semantic Web and Linked Data technologies. In order to learn more about possible ways to overcome this obstacle, we investigated real-life use cases from natural science departments of our university. We interacted with scientists at our university, and developed a first methodology targeted at lowering the barrier for scientists to release their research data as Linked Data.

Our long-term goal is to develop and validate a methodology with appropriate tools support that facilitates the task of publishing research data as Linked Data as well as to assess and compare the cost, feasibility and ease of use of different approaches systematically. In this paper we describe two use cases we are currently implementing. Using these as a springboard we will explore the promises and possible pitfalls of publishing scientific research as Linked Open Data.

2 Use Cases

The main objective of researchers is to provide answers to open scientific questions in their field, thus advancing their own understanding of key problems and phenomena as well as the one of their research field as a whole. Taking on the additional workload of semantically annotating research data will only be

considered if it does not put too much strain on their time budget and can be integrated with their research work. To overcome this obstacle we decided to investigate an approach of co-operation and support. We contacted scientists who are willing to share their data and offered to take care of the technical side of the publication of research data as Linked Data while the scientists contribute their domain knowledge.

We selected two current research projects carried out by natural science departments from Bielefeld University: one from chemistry and one from biology. Both topics are highly interdisciplinary and produce research data that is potentially relevant to researchers from other disciplines. Both scientists were open to the idea of sharing their research data and were willing to contribute their domain knowledge. In the following we will present these two use cases as well as the involved scientists in more detail.

2.1 Chemistry: Glass Transition of Atmospheric Aerosols

Thomas Koop is a professor of Physical Chemistry at Bielefeld University (Germany). He is co-founder and executive editor of the open access journal *Atmospheric Chemistry and Physics*[1]. His research interests include the properties of atmospheric aerosols and their influence on cloud formation. In September 2011, he published a paper on the glass transition of organic aerosols [15].

Aerosols, which consist of floating particles in the air, are an important factor in many atmospheric processes, like light scattering and cloud formation. According to new insights, water soluble organics can form amorphous solids (glasses) in the upper troposphere (i.e. at 8-15 km height), which inhibit ice crystal formation, thereby affecting cirrus cloud formation [18].

In order to quantify the magnitude of this effect, data about the *glass transition temperature* T_g of various substances known to be present in the atmosphere is needed. Because glass transition temperatures are not collected in chemical databases, Thomas Koop conducted an extensive, manual literature research, which took about 100 hours of work. He collected the resulting 596 T_g values from 22 publications in a large spreadsheet-table and supplemented them by additional information like provenance, measurement methods and additional comments.

The corresponding publication [15] does not publish the full list but results aggregated from this data in the form of digrams (an example is shown in Figure 1). A publication of the full dataset as Linked Data, enriched by a semantic representation of the supplementary data, could be very helpful for other scientists and prevent duplication of effort.

Use Case: As a use case we take the example of a researcher in chemistry who wants to collect glass transition temperatures of aerosols. Instead of compiling the data manually from published research articles as Thomas Koop did, our scientist would use Semantic Web search engines to collect relevant data and use

[1] `http://atmos-chem-phys.net/`

appropriate SPARQL queries to aggregate results as needed. Issues that need to be paid attention to are provenance, data quality as well as the fact that different vocabularies might have been used in publishing the data, so that vocabulary harmonization is a crucial part of the process.
Some sample competency questions a researcher could pose to the dataset are:

```
Q: Give me all glass transition temperatures of organic compounds!
Q: Give me all glass transition temperatures of amino acids
   measured by differential scanning calorimetry!
Q: Which substances form glasses at temperature and
   pressure conditions in the troposphere?
```

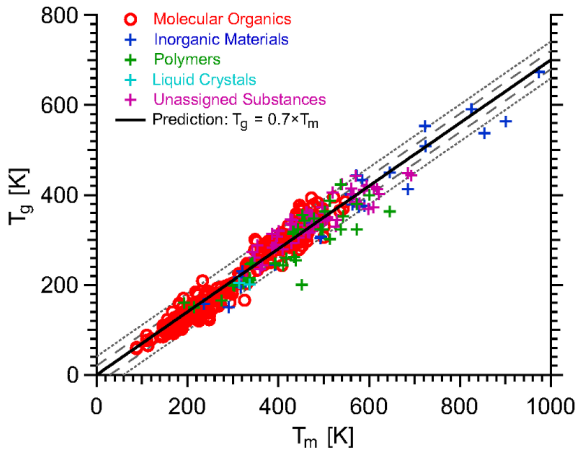


Fig. 1: Graph of the evaluated dataset of glass transition temperatures plotted against melting temperatures. (From [15] - Reproduced with permission of The Royal Society of Chemistry)

2.2 Biology: Natural Movement of Stick Insects

Volker Dürr is a professor of Biological Cybernetics at Bielefeld University (Germany). His research interests include the question of how insects adapt their locomotion behaviour to the context of the situation. He coordinates the EU project EMICAB[2], which has the objective to develop an autonomous hexapod robot.

Insects like the stick insect (Figure 2) can walk on rough terrain, climb obstacles, and use their legs for other behavioural tasks such as searching or reaching [7]. These complex movements are coordinated by a fairly small, experimentally amenable and reasonably well-studied nervous system [5]. Because of the

[2] `http://emicab.eu/`

resource-efficient information processing for solving complex behavioural tasks, the analysis and modelling of insect locomotion have been proposed as a basis for improving artificial autonomous walking robots [10].

The movement of stick insects can be measured by marker-based motion capturing: markers are attached to the body of the insect and tracked by an infrared camera system. The resulting trajectories (time-ordered *xyz*-coordinates) describe the movement of the insect in space. Volker Dürr's group recorded several hours of locomotion sequences from different stick insect species by motion capture. The interpretation of the trajectory data is dependent on the body morphology and the position of the markers on the body. Motion capture datasets have been released in the past, but without specifying the anatomy of the test subject and the exact marker locations, such that these datasets are of limited use outside their original purpose.

A novel approach is to provide sufficient annotation for calculating joint angle time courses for all degrees of freedom from the trajectory data. This would allow the data to be interpreted and reused in other contexts. Pioneering this approach, the EU project EMICAB will make such calculated data publicly available alongside the experimental raw data and metadata about the experimental conditions under which the data was obtained. A semantic annotation of these datasets would greatly improve their retrieval and interpretation by potential future users.

Use Case: A researcher interested in insect motion might download this dataset and extract or recompute the joint angle time courses for all degrees of freedom, thus being able to simulate the organism or compare it to his own results for other or the same organism. The challenge is to incorporate enough information in the data about how the joint angle time courses have been computed so that the comparison is meaningful.
Competency questions the dataset has to answer include:

```
Q: Give me all motion capture datasets about insects!
Q: How large is the complete dataset?
Q: What data is necessary to reproduce the experiment?
```

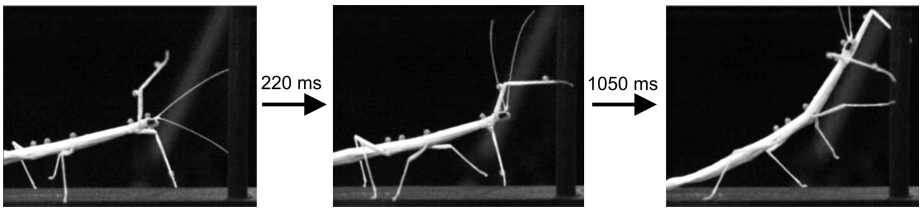


Fig. 2: Stick insect movement with markers for motion capture attached. (Reproduced with permission of Volker Dürr)

3 Methodology

A key element in our methodology is the role of the *scientific data curator*. The data curator's task is to translate the methods and results of scientific research into Linked Data. His role could be compared to that of an investigative reporter: he is not an expert in the domain he is describing, but he is proficient at finding out what is essential and relevant. He asks the scientist the right questions to find out what others need to know to understand and reuse the data. Further, he should be proficient in Semantic Web and Linked Data technologies.

Transferring scientific research into Linked Data can be viewed as a project which requires a joint effort between the scientist and the data curator, such that a close co-operation and constant feedback is essential during all phases of the project. We propose a methodology which involves seven consecutive tasks:

Task 1: Kick-off Meeting
The kick-off meeting is the first meeting of the scientist and the data curator and marks the start of the project. In the kick-off meeting the project members get to know each other and lay the groundwork for the future co-operation. The data curator interviews the scientist about his research interests and gives an introduction into the technology of Linked Data. Ideas and expectations are exchanged in order to build a common understanding of the goal and scope of the project, which will be defined in the next step.

Task 2: Goal Definition
Following the kick-off meeting, the data curator formulates a proposal for the goal and the scope of the project and subsequently refines it by feedback from the scientist. As a main tool at this stage of the project we formulate competency questions, which can be used as tests to make sure that the data contains all relevant information, and to choose vocabularies to represent the data.

For our use cases the goal is to capture all *relevant* data, i.e. the experimental results and all information necessary to reproduce these results. A more lightweight approach could concentrate only on the data *essential* for interpreting the experimental results. The most comprehensive scope would be to use all *available* data, even the pieces that seem irrelevant for the reproducibility of the experiment - but could prove relevant in the future or in other contexts.

Task 3: Knowledge Acquisition
The data curator acquires domain specific knowledge. He achieves this by interviews with the scientists and reading the papers which are based on the experiments. His aim is not to become an expert himself but to get an overview and basic understanding in a short period of time. In addition he collects data which might already be available in structured or semi-structured form.

The glass transition temperatures were collected in a large spreadsheet table with informal comments and undocumented color-coding. The stick insect movement was available in a relational database.

Task 4: Ontology + LOD Exploration
At this stage, the data curator explores existing vocabularies and ontologies that could be reused. He has to thoroughly investigate them in order to evaluate their applicability and usefulness for his task. In addition he is looking for existing Linked Open Data (LOD) datasets that can be linked to. Interlinking and reuse is extremely important, because most of the usefulness of the data lies in its connection to external data. If concepts or resources are involved for which no existing vocabularies or datasets can be located, the data curator will create them. The understanding, evaluation, disambiguation and alignment of existing ontologies is the most important and labour intensive task of the whole process because many existing ontologies are not properly documented.

For our use case in chemistry, several ontologies for the domain of chemistry exist, e.g. CHEMINF [13], ChemAxiom [1] or ChEBI [9]. They differ substantially in scope and complexity. For our use case in biology the *Shape Acquisition and Processing* (SAP) ontology [8] is relevant, which covers the domain of movement data, forms, and virtual characters. The first dataset to look for possible links is *DBpedia*[3], which offers a wealth of concepts and is well dereferenceable for human readers.

Task 5: Implementation
If one or more ontologies have been selected, the data is encoded using the technological tool most appropriate. This could range from the mapping of an existing database, using annotation software or even manual encoding.
Figure 3 presents sample RDF code for encoding a glass transition temperature.

```
:PinicAcid a :ChemicalSubstance ,
  :hasCASNumber "[473-73-4]" ;
  :hasName "Pinic Acid"@en ;
  :hasProperty
    [ a :GlassFormationPoint ;
      dc:source "http://dx.doi.org/10.1039/C1CP22617G" ;
      :hasValue "268.1"^^xsd:float ;
      :hasUnit :Kelvin ;
      :hasStandardDeviation "4.8"^^xsd:float ;
      :hasMeasurementCondition
        [ a :MeasurementPressure ;
          :hasValue "101325"^^xsd:float ;
          :hasUnit :Pascal
        ] ;
      :hasExperimentalTechnique :differentialScanningCalorimtery
    ] .
```

Fig. 3: RDF-representation of the glass transition temperature of pinic acid.

[3] http://dbpedia.org

Task 6: Publication
The data is published, either by uploading the code representing the knowledge to a web server or by importing it into a triplestore. This task essentially completes the project. A SPARQL endpoint should be also provided ideally so that the data can be queried flexibly as needed.

Task 7: Monitoring
Subsequently, the usage of the published data is continuously monitored, for example by looking at SPARQL queries generated by third parties. This can help to improve and refine the data selection and implementation.

Parallel Task: Documentation
Documenting is not a separate task but is done parallel to the other tasks. Like in all projects, documentation plays an important role in forming a common understanding between project members, to proceed from one task to the next, and to enable others to understand and continue the work in the future.

The individual tasks are not strictly linear but feedback loops to earlier tasks are possible if a subsequent task should require correction or refinements to an earlier task. Figure 4 shows an overview of the proposed methodology and possible feedback loops between the tasks.

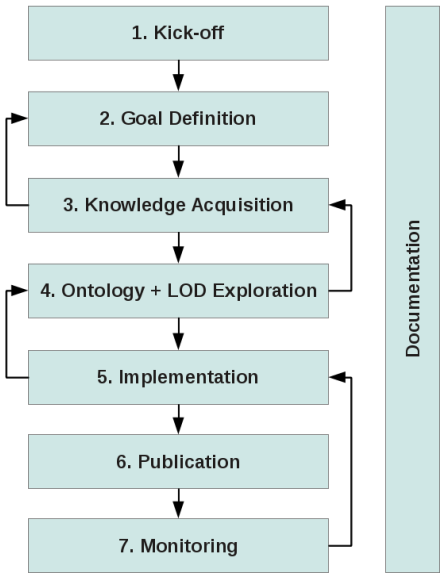


Fig. 4: Workflow for the semantification of research data.

4 Related Work

The *Open Science* movement aims to make scientific publications and research data publicly available. Numerous initiatives have formed over the last few years that put these ideals into practice. Open access journals create alternatives to the old publication system, e.g. *Atmospheric Chemistry and Physics*[4], which has been using an open review system for 10 years and has the highest impact factor of all 68 journals in the field of meteorology and atmospheric sciences. Even traditional publishers are beginning to embrace the new technologies, like Elsevier did with its *Grand Challenge*[5]. Universities are setting up public repositories of research data, e.g. VIVO[6] or *Potsdam Mind Research Repository*[7], which gives access to peer-reviewed publications and additional data and scripts for analyses and figures. Large Datasets have been opened, like the *Human Genome Project*[8] or the *Sloan Digital Sky Survey*[9]. The W3C's *Health Care and Life Sciences Interest Group* (HCLSIG)[10] created a knowledge base of RDF data from the domains of health care and the life sciences. Social networks like myExperiment[11] allow scientists to publish and share their scientific workflows. With all of these the ideas of Open Science are gradually changing the way scientific research is done.

One of the main tasks of our methodology is the elicitation of knowledge from the domain experts. Methodologies for knowledge extraction have a long tradition in knowledge management (cf. [14]). Especially relevant to our approach is the work on ontology engineering [17]. A data curator does not have the primary goal of creating an ontology or vocabulary, but he may find it necessary to develop a vocabulary, or extend an existing one, if no existing ontology for a specific task can be found. In any case, he needs a good understanding of methodologies for the creation and evolution of ontologies, in order to evaluate and apply them.

For an efficient creation of Linked Data several approaches have been developed, either by automated translation or by tool-support for the author. Four kinds of approaches can be distinguished:

1. Export from existing sets of structured data: relational databases, like the ChEBI database[12], which collects data about chemical substances, are exported into Linked Data by mapping database fields to a vocabulary. The D2R Project[13] exposes the content of a relational database as Linked Data.

[4] http://atmos-chem-phys.net/
[5] http://www.elseviergrandchallenge.com/
[6] http://vivoweb.org/
[7] http://read.psych.uni-potsdam.de/pmr2/
[8] http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
[9] http://www.sdss.org/
[10] http://www.w3.org/wiki/HCLSIG
[11] http://www.myexperiment.org/
[12] http://www.ebi.ac.uk/chebi/
[13] http://d2rq.org

2. Export from content management systems (CMS): Drupal[14], WordPress[15] can publish editorial content as Linked Data using pre-selected vocabularies.
3. Automated extraction of data from scientific publications by text-mining techniques: Several methods for automatic extraction of bibliographic metadata have been developed [12]. The OSCAR3 [6] programme identifies chemical terms by natural language processing.
4. Semantic annotation of papers either by editors or by scientists: within the *Prospect* project[16] for instance, the Royal Society of Chemistry (RSC) has taken the approach to have papers semantically enriched not by the scientist but by editors (cf. [3] for an overview).

Which of these approaches is the best one for a specific task depends on the goal and the scope of the individual project. Because our goal is to develop deeper insights into how existing vocabularies and ontologies can be reused in the process of publishing Linked Data, we have decided to carefully evaluate and select the most appropriate vocabularies instead of converting the data to RDF using some automatic approach (e.g. RDB2RDF[17]). In the future we plan to compare the results with those of more automatic approaches.

5 Conclusion and Future Work

We presented two use cases from chemistry and biology that we are currently working on. The specific aim of these projects is to publish the relevant scientific research data as Linked Data, i.e. the results of the experiments and the experimental set-ups necessary to reproduce the results. We proposed a methodology that is characterized by a close co-operation between a scientist and a scientific data curator, who translates the scientists' domain knowledge into Linked Data. This preliminary methodology will be validated and refined empirically as the implementation progresses.

In preparing the projects we found that all scientists we interacted with are interested in the ideas and possibilities of Linked Open Data. But only few of them are willing to contribute and invest their data, time and knowledge. A close co-operation between scientist and data curator is highly important for the success of the project. Therefore trust is essential. The scientist must be sure that his data is handled responsibly and that his wishes regarding its publication are respected.

So far we have defined the goals and the scope of both projects and elicited the relevant domain knowledge. We are currently in the process of evaluating suitable existing ontologies and Linked Data from other datasets we could link to. Because our goal to publish all relevant research data is rather ambitious,

[14] `http://drupal.org/`
[15] `http://wordpress.org`
[16] `http://www.rsc.org/Publishing/Journals/ProjectProspect/`
[17] `http://www.w3.org/2001/sw/rdb2rdf/`

the cost involved with each of the steps is high. Especially the exploration and evaluation of existing ontologies has proven to be complex and time consuming.

Our long-term objective is to contribute to the formation of an open research infrastructure by empowering scientists to publish their research as Linked Data. Towards this goal, appropriate methodologies for the transformation of research data into Linked Data are needed. In combination with shared ontologies and tool support, we expect these to be the foundation for scientists to adopt the new technology of Linked Data. Our hypothesis is that the role of a *scientific data curator* as proposed in this paper is a key function towards facilitating this development.

After completing the two use cases we will perform a thorough analysis of the resulting datasets and of the overall process. Focus will be put on the question of the cost involved for each of the tasks. Our next step will be the development of criteria for choosing, combining and expanding existing ontologies. In future work we plan to use our manually created Linked Data as a gold standard for the evaluation of less expensive, semi-automatic or fully automatic solutions.

Acknowledgements

We are grateful to Thomas Koop and Volker Dürr for sharing their research data and their helpful insights.
This work is funded as part of the Center of Excellence Cognitive Interaction Technology (CITEC) at Bielefeld University.

References

1. Adams, N., Cannon, E., Murray-Rust, P.: Chemaxiom - an ontological framework for chemistry in science. Available from Nature Precedings: `http://dx.doi.org/10.1038/npre.2009.3714.1` (2009)
2. Alliance of German Science Organisations: Priority Initiative "Digital Information". Retrieved April 12, 2012, from: `http://www.wissenschaftsrat.de/download/archiv/Allianz-digitale%20Info_engl.pdf` (June 11, 2008)
3. Attwood, T.K., Kell, D.B., McDermott, P., Marsh, J., Pettifer, S.R., Thorne, D.: Calling international rescue: knowledge lost in literature and data landslide! Biochemical Journal **424**(3) (2009) 317–333
4. Berners-Lee, T., Fensel, D., Hendler, J.A., Lieberman, H., Wahlster, W., eds.: Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential. MIT Press, Cambridge, MA (2005)
5. Büschges, A., Akay, T., Gabriel, J.P., Schmidt, J.: Organizing network action for locomotion: Insights from studying insect walking. Brain Res. Rev. **57**(1) (January 2008) 162–171
6. Corbett, P., Murray-Rust, P. In: High-throughput identification of chemistry in life science texts. Volume 4216. Springer Berlin Heidelberg (2006) 107–118
7. Cruse, H., Dürr, V., Schilling, M., Schmitz, J.: Principles of insect locomotion. In Arena, P., Patanè, L., eds.: Spatial temporal patterns for action-oriented perception in roving robots. Springer, Berlin (2009) 43–96

8. De Floriani, L., Hui, A., Papaleo, L., Huang, M., Hendler, J.: A semantic web environment for digital shapes understanding. In: Proceedings of the semantic and digital media technologies 2nd international conference on Semantic Multimedia. SAMT'07, Berlin, Heidelberg, Springer-Verlag (2007) 226–239
9. Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., Ashburner, M.: Chebi: a database and ontology for chemical entities of biological interest. Nucleic Acids Research **36**(suppl 1) (2008) D344–D350
10. Dürr, V., Schmitz, J., Cruse, H.: Behaviour-based modelling of hexapod locomotion: linking biology and technical application. Arthropod Struct Dev **33**(3) (2004) 237–50
11. Feijen, M.: What researchers want - a literature study of researchers' requirements with respect to storage and access to research data. Retrieved April 12, 2012, from SURFfoundation: `http://www.surffoundation.nl/nl/publicaties/Documents/What_researchers_want.pdf` (February 2011)
12. Groza, T., Grimnes, G., Handschuh, S., Decker, S.: From raw publications to linked data. Knowledge and Information Systems 1–21
13. Hastings, J., Chepelev, L., Willighagen, E., Adams, N., Steinbeck, C., Dumontier, M.: The chemical information ontology: Provenance and disambiguation for chemical data on the biological semantic web. PLoS ONE **6**(10), DOI 10.1371/journal.pone.0025513: `http://dx.doi.org/10.1371%2Fjournal.pone.0025513` (10 2011)
14. Holsapple, C., ed.: Handbook on knowledge management. International handbooks on information systems. Springer, Berlin (2003)
15. Koop, T., Bookhold, J., Shiraiwa, M., Pöschl, U.: Glass transition and phase state of organic compounds: dependency on molecular properties and implications for secondary organic aerosols in the atmosphere. Phys. Chem. Chem. Phys. **13** (2011) 19238–19255
16. Popper, K.R.: The Logic of Scientific Discovery. Hutchinson, London (1959)
17. Sure, Y., Staab, S., Studer, R.: Ontology Engineering Methodology Handbook on Ontologies. In Staab, S., Studer, R., eds.: Handbook on Ontologies. International Handbooks on Information Systems. Springer, Berlin, Heidelberg (2009) 135–152
18. Zobrist, B., Marcolli, C., Pedernera, D.A., Koop, T.: Do atmospheric aerosols form glasses? Atmos. Chem. Phys. **8**(17) (2008) 5221–5244

Automated Assembly of Custom Narratives from Modular Content using Semantic Representations of Real-world Domains and Audiences

Joshua Wulf, David Jorm, Matthew Casperson, and Lee Newson
jwulf,djorm,mcaspers,lnewson@redhat.com

Red Hat Engineering Content Services

Abstract. We present an approach to automatically assembling customized technical documents covering a specified area of interest, tailored to the needs of a specific audience, and with a meaningful narrative structure using semantically annotated modular units of information (topics), and ontologies that describe the structure of the real-world domain of interest. In this paper we explore the nature of narrative, and how an automated document assembler can produce coherent narrative using semantic representation. We introduce a Semantic Publishing system, named Skynet, that implements these ideas in the context of documenting commercial software products.
Keywords: semantic publication, automated assembly, customized narrative

1 Introduction

A utomated Document Assembly is the aggregation by a software agent of smaller units of information into a larger structure to meet the information requirements of a specific audience.

We define the larger structure as a *document*, which may be instantiated as a static linear document on paper (or as a pdf), or as a customized view of hyperlinked text. Narrowing the definition used by Andre et al. [AFQ], we attempt to formally define a document as "a collection of information targeted to an audience interest, constrained to include relevant information and exclude irrelevant information, and grouped and sequenced to match the audience's hierarchy of concern".

We distinguish technical documents – documentation that seeks to inform the reader about factual information – from other types of documents such as prose, or fiction, which fall outside our definition of a document (see Wright [Wri]).

The optimum structure for a document is a function of audience range of interest, existing audience knowledge, and audience hierarchy of concern. To produce an optimum document structure, an automated document assembler must have knowledge of these three aspects of the audience. Additionally it

requires knowledge of the domain within the audience's interest, a collection of modular units of documentation that describe this domain, and knowledge of how these units relate to the domain and each other.

This allows an automated assembler to produce a customized document describing the range of the domain that matches the audience's range of interest, in a way that matches the audience's hierarchy of concern and level of knowledge. This can be done in the form of bespoke pdf documents or by modifying the dynamic presentation of a hyperlinked web of information.

We will look at each of these four areas in turn: Semantic representation of the domain; Modular Units of Content; Semantic Representation of an Audience; Automated Custom Assembly.

Before examining these four areas, however, we explain the theory of cognition that informs our implementation.

2 Theory of Cognition

We use a model of communication based on the Theory of Cognition proposed by van Dijk and Kintsch [DK]. They propose a categorization of communication elements that includes *textual representation* - words used to describe something - and a *situation model* - an internal mental model. The situation model represents some real-world domain, while the textual model represents a situation model in language.

The process of technical communication in this model is one of deconstructing the internal mental model possessed by a subject matter expert, marshaling the elements of that mental model into a verbal or written representation, streaming that representation to a receiver, demarshalling those verbal or written representations into mental elements in the mind of the receiver, and integrating those elements into the receiver's internal mental model.

All three of: the textual representation; the situational model; and a mapping between the two, must be available to the automated processor. We will first examine the situational model, which is made available to the automated processor as a *semantic representation of a real-world domain.*

3 Semantic Representation of a Real-World Domain

A situation model is an internal predictive mental model that is used to predict how objects in the real-world will act and react. The situation model encodes categories, membership, and relationship between elements of the world of experience [Gar] [Joh].

The structure of a document is itself semantic - the spatial and temporal relationships between textual units convey information about the relationships between the real-world elements that the textual units represent. Important information appears before less important information; things that occur or are encountered first are presented first; dependencies are presented before the things

that depend on them. These kinds of decisions about the structure of a document are made by a human author using their own internal mental model. In our experience, in cases where a human author is missing or has an incomplete mental model of a domain, they must take recourse to subject matter experts for guidance on where to place information (for an illustrative example, see this discussion).

We define a *coherent narrative* as a semantic structure composed of comprehensible textual units. In our system textual units are authored by human authors, and their assembly into a meaningful (semantic) structure, or coherent narrative, is the role of the automated processor.

We conceptualize a situation model as an n-dimension hypercube which encodes a multiplicity of relationships along different dimensions of interest. This n-dimensional hypercube enables us to formally (and programatically) answer the question: "*Why is it that certain sentences should be "close" to each other in an instructional document ... ?*"[Hor] This Semantic Representation of a Real-World Domain (Domain Model) acts as a situation model for a automated document assembler that performs the role of human author/subject matter expert in deciding what information to include/exclude, and how to structure it in response to a given audience.

The Domain Model is implemented as categories and tags, which represent dimensions and points, respectively, in the n-dimensional hypercube of the situation model that the Domain Model simulates.

3.1 Ordered and Unordered Dimensions

Construction of a coherent narrative involves grouping and sequencing operations. Related information is grouped into sections and chapters, sequenced within that container, and the containers are themselves are grouped and sequenced.

We distinguish between *ordered* and *unordered* dimensions in the Domain Model, to encode the bases for these grouping and sequencing decisions.

Ordered dimensions are those dimensions whose points have an intrinsic sequential relationship that should be considered when constructing a narrative. Examples of such dimensions include "Lifecycle" (which is tied to dependency and also to time), "Temperature", and "Location" (for example: layers in a software stack).

Unordered dimensions are those dimensions whose points belong to the dimension, but do not have an intrinsic sequential relationship. Examples of such dimensions include: "End User Demographic", and "Name".

Unordered dimensions cannot be used as the basis for sequencing operations, and when information must be sequenced on the basis of a common unordered dimension the correct convention to use is "alphabetical ordering", a structure that semantically communicates: "*There is no meaning to this ordering*" [1].

[1] It is important to note for implementation purposes that alphabetical ordering is completely extrinsic to the semantic dimension, as it will change in the document output depending on the target language.

Table 1. Example Ordered Dimension in a Domain Model

Category	Tag	Tag Order
Lifecycle	Prerequisites	0
	Download	1
	Installation	2
	Configuration	3
	Deployment	4
	Shut down	5
	Redeployment	6
	Upgrade	7
	Removal	8

Table 2. Example Unordered Dimension in a Domain Model

Category	Tag	Tag Order
End User Concern	Application Development	null
	Server Administration	null
	Migration	null
	Troubleshooting	null

Tags belong to categories, and categories that encode ordered dimensions (*Category.IsOrdered*) make use of the *TagOrder* property of tags to encode the ordered nature of the dimension.

4 Modular Units of Content

In order to construct a document, an automated assembler requires a semantic representation of a situation model, and modular units of content to assemble.

The modular units must be sufficiently atomic to be meaningfully mapped to points within the n-dimensional hypercube of the Domain Model. A modular unit may be mapped to multiple points within the Domain Model, but all of the content in the unit must map to the same point(s). Otherwise, inclusion of the content in the output document based on its mapping to the Domain Model will result in the inclusion of content "in the wrong place".

To achieve this level of atomicity we have adopted use of the Darwin Information Typing Architecture (DITA) [DITA] topic types. The DITA Topic Types are based in part on the Information Mapping work of Horn [Hor+1]. Documents produced using formal division into Information Mapping units have been shown to be more effective than those produced with an ad-hoc information architecture [Hor+2].

The information in a DITA topic is constrained to a single subject, and a single information role. This level of atomicity makes them ideal candidates for mapping to a Domain Model through metadata tagging. We implement the modular units using the DITA categorizations of concept, task, and reference topics, but using the Docbook XML schema, to leverage our existing open-source Docbook publishing toolchain, Publican. It is the information typing aspect of DITA [IMI] that is of principal use and interest to us[1].

Constraining the content of a textual unit (topic) to a single *subject* means that they can be unambiguously mapped to points inside the n-dimensional hypercube of the Domain Model. This allows them to be reliably assembled according to the macrostructure of the document. Constraining them to a single *information role* means that they can be reliably assembled within that macrostructure, as concepts, references, and tasks play deterministic roles in the textual representation of an area of a situation model (something we will examine in due course). Some additional information may be required to assemble the units into a coherent narrative at this level. In addition to mapping topics to the Domain Model, topics can be mapped to each other. So a task can declare that a specific concept "is a dependency", or a reference can declare that it "illustrates" a specific task.

4.1 Natural Language Processing and Ontology Tagging

The content of the textual units (topics) is generated by human authors. The metadata tagging of the topics against the Domain Model and against each other is also performed by human authors. We have some rudimentary natural language processing tools that assist in this process. A conceptual vocabulary is generated based on the *Title* property of the topic. A scanner then examines the textual content of other topics, and suggests potential relationships based on the content, which can then be accepted or rejected by a human moderator. This is similar to the approach used by the BBC World Cup semantic publishing system [BBC].

We further examine the role that topic information role plays in assembling output when we examine Automated Custom Assembly.

5 Semantic Representation of an Audience

Generally speaking, an audience is a group of people with a shared interest and level of pre-existing knowledge [MS]. Audience is usually an approximation of a range, within which individual readers may completely or partially fit. The economics of document production dictate that a small number of documents be produced to serve a large number of people, and hence the idea of "audience" as a range. With automated assembly and electronic delivery, however, the economics of production of narrative change, and the problems associated with defining an

[1] although we also plan to support DITA XML encoded content in the future.

audience as generalized ranges [EL] [Ong] can be mitigated by using very specific definitions.

Different people are interested in different information, and they are interested in the same information in different ways.

Consider the following two cases:

1. An organization where the layers of a software stack are horizontally divided. In this case, one group is reponsible for the database component, and another group is responsible for the Operating System component.
2. An organization where the layers of a software stack are vertically divided. In this case, one group is responsible for installing both the Operating System and the Database, while another group is responsible for maintaining them.

In these two distinct cases the same information is needed by people in either organization, but it is needed in a different combination in each case.

Customized Narratives can be generated for each of these use cases. In the first case the narrative is generated by creating two documents. The information in the first document is constrained to topics tagged with the "Component:Database" tag. The information in the second document is constrained to topics tagged with the "Component:Operating System" tag.

In the second case, two documents are created. Both documents are constrained to ("Component: Database" OR "Component: Operating System"). The content of the documents is also grouped at the first level on these tags, resulting in two sections: "Database" and "Operating System". If the "Component" category is implemented as an ordered category based on the software layer, then the Operating System section will precede the Database section, otherwise they will be alphabetically ordered.

The first document is further constrained to information tagged "Lifecycle: Installation", and the second document is constrained to information (NOT tagged "Lifecycle: Installation").

Audiences can be further defined by linking an audience with a list of concepts that they can be expected to know. These concepts can then be elided from documents that are produced for this audience. Tasks can also be tagged with a tag from an ordered category "difficulty", and a threshold set for an audience, so that introductory and advanced guides can be produced.

When a formal definition of an audience is available to an automated processor, it can use this definition in conjunction with a semantic representation of a situation model and semantically-annotated modular units of content to assemble a custom narrative relevant to the audience's needs.

Because the production cost of this narrative assembly is so low, documents produced for an audience of one become economically feasible. The costs of narrative production move away from human authors (who can hardly be expected to write a different book for each reader), and move to processor cycles and a cognitive cost on readers, which we examine later.

6 Automated Custom Assembly

We use an algorithm to solve the general case of creating a coherent narrative structure from an arbitrary collection of textual units:

To automatically generate a semantic structure, we examine all of the topics that have been returned for a query, and assemble them into intermediate units based on topic types and any declared relationships between topics. We then examine the resulting aggregate units to determine if there exist meta data dimensions in which we can locate all of the units in the query.

1. If there is no meta data category for which all topics at this order of structure have a meta data tag, it's alphabetical ordering.
2. If all topics at this order of structure have a meta data tag from the same sequenced category, that category is a candidate for sequencing.
3. If all topics at this order of structure have a meta data tag from the same non-sequenced category, that category is a candidate for grouping.
4. If one grouping candidate exists and no sequencing candidates exist, group on that category and sequence alphabetically.
5. If one sequencing candidate exists and no grouping candidates exist, group and sequence on that category.
6. If more than one grouping or sequencing candidate exists, then follow the semantic rules for hierarchy of concern.

In addition to the interaction of audience range of interest and hierarchy of concern with the semantic representation of the situation model, the information type of the textual units (topics) influences the output structure of the document.

We use some basic patterns to structure the output, all other semantic considerations being equal, based on topic type. The basic pattern we use is "Concept, Task, Reference". Relevant or dependent concepts precede a Task, which is followed by additional reference material, including any example that illustrates its use.

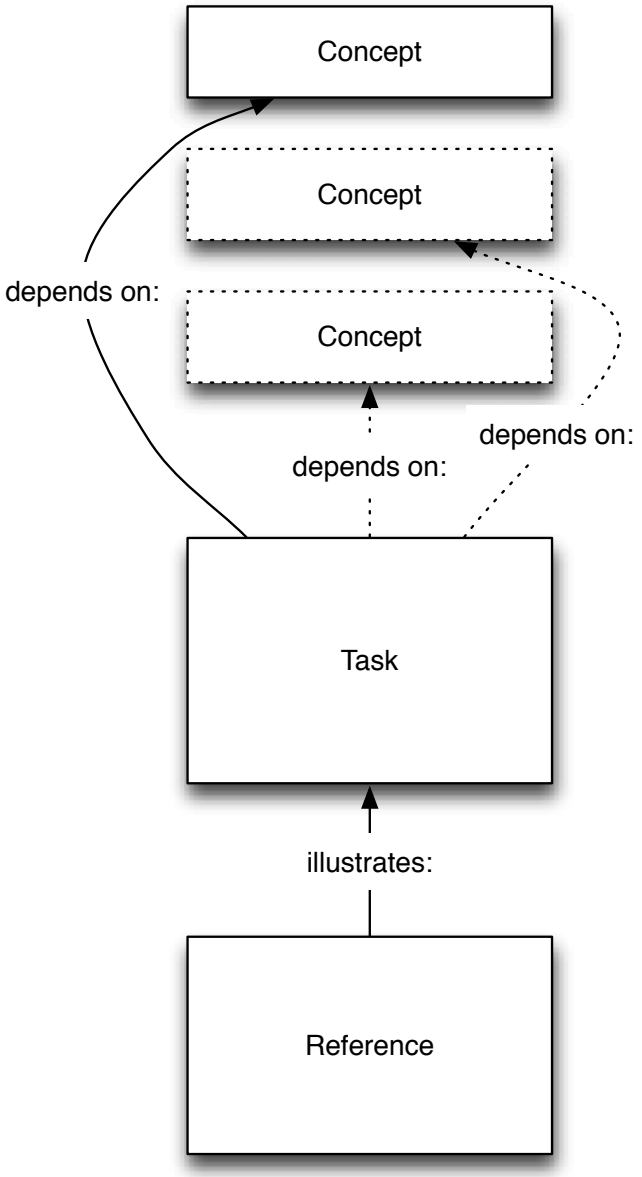
Figure 1 illustrates the output structure of a group of related topics, based on their topic type. The dotted lines represent concepts that may be elided or collapsed (in an html output) depending on the audience's level of knowledge, or the predicted relevance of the concept.

7 Current Status of Our System

Currently our semantic publishing system, known internally as Skynet and under heavy development, is implemented using as a JBoss Seam application, with a MySQL database to store the topics, and a topics-to-tags, tags-to-categories database schema to implement extensible meta data. Our processing engine is implemented as a combination of procedural code and rules using JBoss Drools.

The system is implemented as a platform, and has a REST API interface that allows it to be easily integrated and extended. One of the first extensions that we've developed for it is a content specification processor that allows an arbitrary

Fig. 1. An assembly pattern for textual content based on topic type



topic map to be passed to the system and returned as a Docbook book. This allows external semantic processors to generate their own output structures, and request the platform to build it from the content in the repository. This open and extensible design allows us to innovate and extend outside of the core code of the project.

The system is under development as an open source project on Sourceforge, and the current implementation is running on the Red Hat internal network and being used to develop the product documentation for the upcoming release of JBoss Enterprise Application Platform 6.

The documentation for JBoss Enterprise Application Platform 6 is written as modular content units (topics), and tagged against a semantic representation. The final output documentation is generated by an automated processor that locates the modular content units in an aggregate structure using the semantic representation to generate the document structure. In this sense it functions in much the same way as the BBC World Cup semantic publishing system [BBC] — the automated processor handles the publishing into the larger structure, while the human author is concerned with writing the content, and accurately describing it in terms of the semantic representation.

The document structure in Figure 2, for the JBoss Enterprise Application Platform 6 documentation, is generated by an automated processor using our semantic representation and semantically-annotated textual units. Information is constrained to the area of the hypercube containing "JBEAP 6" tagged topics. It is then grouped at the first level on the "Technology Component" dimension, which is an unordered dimension, hence the alphabetical ordering. At the second level of structure it is grouped on "End-User Concern", which is an ordered dimension based on lifecycle.

This is shown here as a tree structure, but it could also be instantiated as the table of contents of pdf output.

8 Challenges and Opportunities

At this point in time we are able to generate multiple output structures for various audience definitions. We have the data in place to allow the generation of Active Documents [DGK], however our production infrastructure currently serves over 2TB of documentation to users each month, so introducing dynamic content generation on our public-facing website represents a scalability and security challenge.

Our next plan is to make our dataset available to the public, possibly through a web service end-point, with our semantic representation available as RDF data. This will allow users to design their own documentation. In this case, users will be able to express their interest, either at high level or as a specific, complex query, and we will return a pdf file.

A significant challenge is encapsulating the complexity of the system. Users are not used to defining the content and architecture of documents. Making

X

Fig. 2. Output structure created by an automated processor using a semantic representation of a real-world domain and semantically-annotated textual units

the power of the system available to users without overwhelming them with its complexity is the greatest challenge and opportunity.

We are investigating three avenues to work towards this: natural language question answering; progressive customization of view; ant colony optimization of the semantic representation.

8.1 Natural Language Question Answering

Users are not accustomed to designing books, or articulating the exact nature of their domain of interest in a wide sense. They are, however, used to searching for information relating to a specific query that they have. With a semantic representation and semantically-annotated textual units, we are now in a position to investigate generating customised answers to queries from units of documentation. Rather than attempting to build a complete book, and requiring a user to specify an entire book, we would attempt to assemble the pieces to answer a specific query. Some disambiguation questions may be necessary to derive the exact area of interest, and the user's preferred format (hierarchy of concern), and then we can produce a small document to answer their query.

8.2 Progressive Customization of View

Rather than requiring users to define a view of the information, we can present users with a default view of the information (as we do now with JBoss Enterprise Application Platform 6 documentation). However, over time we can customize that view based on the user's behavior. We can infer things about the user based on their interaction with the material. When a user clicks on a search result, we know something about the result that they have clicked on. If we detect a preference to one area of the hypercube, we can weight search results to favor that area. We can establish *weak assumptions* about the user based on this kind of behavior. If we introduce the ability for the user to provide us with qualitative feedback, such as a "Like" or "this is what I was looking for" button, then we can also create *strong assumptions* about the user's preferences and further weight customizations.

Progressive customization of view is less taxing on users, although it may be more taxing on hardware requirements. Offline static rendering in response to a specific user query pushes the burden more to the user's side, and will be the first approach for the early adopters.

8.3 Ant Colony Optimization of the Domain Model

Ant Colony Optimization [DD] is a meta-heuristic optimization method, that simulates the behavior of an ant colony to approximate an optimal solution. It relies on many agents, in this case users, to iteratively explore the solution space and approximate a global optimum.

The Domain Model allows us to formally state "*Why is it that certain sentences should be "close" to each other in an instructional document ... ?*"[Hor].

However, there may be dimensions of interest to users that are not captured in our Domain Model, or are incorrectly encoded. If the Domain Model captures the situation model accurately, then rotation of the hypercube should allow two points in the model to come into proximity with each other. This means that information that is required sequentially by the user will be presented sequentially when the user's axis of interest is used to orient the hypercube. If we find that users consistently search for and then "like" two topics within a defined temporal period, and that these two topics are not available in proximity in a rotation of the hypercube, then it is an indication that there is something missing from the Domain Model, and this can be examined.

References

AFQ. Andre, J., Furuta, R.K., Quint, V: "Structured Documents", Cambridge University Press, 1989, pp. 3.

BBC. Rayfield, J., "BBC World Cup 2010 dynamic semantic publishing", BBC Internet Blog, 2010, http://www.bbc.co.uk/blogs/bbcinternet/2010/07/bbc_world_cup_2010_dynamic_sem.html.

DD. Dorigo, M., Di Caro, G. The Ant Colony Meta-Heuristic, IRIDIA Universit Libre de Bruxelles, 1999

DGK. David, C., Ginev, D, Kohlhase, M., Matican, B. Mirea, S., A Framework for Semantic Publishing of Modular Content Objects, In: *Proceedings of the 1st Workshop on Semantic Publishing 2011*, 2011, http://ceur-ws.org/Vol-721/paper-03.pdf.

DITA. OASIS Darwin Information Typing Architecture (DITA) Version 1.2 Specification, OASIS Standard, 2010, http://docs.oasis-open.org/dita/v1.2/spec/DITA1.2-spec.html.

DK. van Dijk, T. A., Kintsch, W.: Strategies of Discourse Comprehension, New York Academic Press, 1983, pp 305.

EL. Ede, L., Lunsford, A. "Audience Addressed/Audience Invoked: The Role of Audience in Composition Theory and Pedagogy", In: *College Composition and Communication*, Vol. 35, No. 2, National Council of Teachers of English, 1984.

Gar. Garnham, A.: "Mental Models As Representations of Discourse and Text", Ellis Horwood Ltd, 1988.

Hor. Horn, R.E., "Structured Writing as a Paradigm", N. J., Educational Technology Publications, 1998.

Hor+1. Horn, R.E., "Information Mapping", In *Training in Business and Industry*, Vol. 11, No.3, March 1974

Hor+2. Horn, R.E., "How High Can It Fly? Examining the Evidence on Information Mapping's Method of High Performance Communication", The Lexington Institute, 1992.

IMI. Information Mapping Institute Whitepaper, "Information Mapping©and DITA: Two Worlds, One Solution", Information Mapping Institute, 2011, http://www.informationmapping.com/us/resources/whitepapers/261-information-mapping-and-dita

Joh. Johnson-Laird, P.N.: "Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness", Cambridge: Cambridge University Press 1983.

MS. Mathes, J.C., Stevenson, Dwight W."Designing Technical Reports: Writing for Audiences in Organizations", The Bobbs-Merrill Company, Inc., 1976.

Ong. Ong, W.J., “The Writer’s Audience is Always a Fiction”, In: *PMLA*, Vol. 90, No. 1, Modern Language Association, 1975

Wri. Wright, P.: “Writing Technical Information”. In: *Review of Research in Education* Vol.14, (1987) pp.327-385 American Educational Research Association.