Proceedings of the 3rd Workshop on Semantic Publishing (SePublica 2013)

10th Extended Semantic Web Conference Montpellier, France, 26 May 2013

edited by Alexander García Castro, Christoph Lange, Phillip Lord, and Robert Stevens

26 May 2013

Preface

This volume contains the full papers presented at SePublica 2013 (`http://sepublica.mywikipaper.org`), the Third International Workshop on Semantic Publishing (Machine-comprehensible Documents, Bridging the Gap between Publications and Data) held on 26 May 2013 in Montpellier, France.

There were 7 submissions. Each submission was reviewed by at least 2, and on the average 3.1, programme committee members. The committee decided to accept all of them.

The programme also included an invited talk given by Peter Murray-Rust from the University of Cambridge, UK, on the question of "How do we make Scholarship Semantic?", which is included as the first paper in this volume.

It also includes five presentations of *polemics*, which are not included in this proceedings volume but archived in the Knowledge Blog at `http://event.knowledgeblog.org/event/sepublica-2013`:

- Hal Warren, Bryan Dennis, Eva Winer: Flash Mob Science, Open Innovation and Semantic Publishing
- Idafen Santana Pérez, Daniel Garijo, Oscar Corcho: Science, Semantic Web and Excuses
- Chris Mavergames: Polemic on Future of Scholarly Publishing/Semantic Publishing
- Sarven Capadisli: Linked Research

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 and ceur-make, which made this task really convenient.

29 May 2013
Birmingham

Christoph Lange
Alexander García Castro
Phillip Lord
Robert Stevens

Program Committee

Paolo Ciccarese	Harvard Medical School & Massachusetts General Hospital
Philippe Cudré-Mauroux	University of Fribourg
Sudeshna Das	Harvard University
Kai Eckert	University of Mannheim
Alexander Garcia	Florida State University Guest Professor
Leyla Jael García Castro	Universität der Bundeswehr
Benjamin Good	TSRI
Tudor Groza	School of ITEE, The University of Queensland
Christoph Lange	University of Birmingham
Phillip Lord	Newcastle University
Robert Morris	DCR Consulting, LLC, UMASS-Boston, and Harvard University
Steve Pettifer	University of Manchester
Jodi Schneider	DERI, NUI Galway
Robert Stevens	University of Manchester
Daniel Vila	
Jun Zhao	University of Oxford

Additional Reviewers

R

Ritze, Dominique

Contents

How do we make Scholarship Semantic?
Peter Murray-Rust 6

Twenty-Five Shades of Greycite: Semantics for referencing and preservation
Phillip Lord and Lindsay Marshall 10

Systematic Reviews as an interface to the web of (trial) data: Using PICO as an ontology for knowledge synthesis in evidence-based healthcare research
Chris Mavergames, Silver Oliver and Lorne Becker 22

Towards Linked Research Data: An Institutional Approach
Cord Wiljes, Najko Jahn, Florian Lier, Thilo Paul-Stueve, Johanna Vompras, Christian Pietsch and Philipp Cimiano 27

Repurposing Benchmark Corpora for Reconstructing Provenance
Sara Magliacane and Paul Groth 39

Connections across scientific publications based on semantic annotations
Leyla Jael García Castro, Rafael Berlanga, Dietrich Rebholz-Schuhmann and Alexander Garcia 51

Towards the automatic identification of the nature of citations
Angelo Di Iorio, Andrea Giovanni Nuzzolese and Silvio Peroni 63

How Reliable is Your workflow: Monitoring Decay in Scholarly Publications
José Manuel Gómez-Pérez, Esteban García-Cuesta, Jun Zhao, Aleix Garrido and José Enrique Ruiz 75

How do we make Scholarship Semantic?

Peter Murray-Rust,

Unilever Centre for Molecular Sciences Informatics, Department of Chemistry, University of Cambridge, CB2 1EW, UK

This paper is an account of an invited keynote to the SePublica Workshop of the ESWC 2013 meeting.

I am grateful for the opportunity to comment on the opportunities and challenges of using semantic approaches in scholarly publishing. ("Scholarship" covers many fields of endeavour and should extend beyond the ivory tower of academia.) We were asked to provide crazy ideas and a polemic, where arguments are backed by personal conviction as much as proven experience. The primary aim of a polemic is to galvanize people into action and the ideas in this paper were aired as several blog posts shortly before the SePublica meeting.

There are very few examples of semantic publishing because the present culture militates against it. If an author sends a semantic manuscript to a journal it will almost certainly be rejected or dumbed down to a more primitive format. To be fair, proper publication requires considerable work from the journal editors and to get semantic benefit, the reader probably has to have special functionality installed. There have been a number of one-off attempts to publish semantically (including some of my own) but they haven't "caught on".

So for most of us, including the readers of this article, semantic publications are an act of faith. We believe them to be valuable, and that when the literature is semantic a brave new world will emerge. I had the privilege of hearing TimBL at CERN at WWW1 in 1994 and he changed my life. The picture of a semantic world mirroring the human and physical world was immediately obvious and imperative. The current problem is that we know it takes a revolution which is not only technical, but cultural, and it is surprising how slow it is proving.

In this polemic I suggest that we limit the level of semantics to simple concepts, (some similarities to TimBL's 5-stars of data):

- We have **to be a community**.
- We have to **identify things** that can be described and on which we are prepared to agree.
- We have to **describe things**
- We have to **name things**
- We have to be able to **find things** (addressing)

All of this is sellable to those who use the web – we don't need formal logic and Aristotelian ontologies. We need identifier systems, ideas of objects and classes, and widely distributed tools. DBpedia and Schema.org are good enough to start with. If this is all that we manage to introduce to scholarly publishing that will be a major success.

So why are semantics important for scholarly publishing? At the most basic level it is about control. The people who control our semantics will control our lives. Semantics constrain the formal language we use and that will constrain our natural language. We humans may not yet be in danger of Orwell's Newspeak but our machines will be. And therefore we have to assert rights to have our say over our machines' semantics.

With a few exceptions I have lost faith in the ability of scholarly societies to act as leads in information development. The problem is that many of them are also major publishers and if they, not us, decide how we are able to express ourselves then it will be based on cost-effectiveness and control, not on what we believe is correct.

The major task for SePublica is to devise a strategy for bottom-up Open semantics. That's what Gene Ontology did for bioscience. We need to identify the common tools and the common sources of semantic material. And it will be slow – it took crystallography 15 years to create their dictionaries and system and although we are

speeding up we'll need several years even when the community is well inclined. Semantics have to be Open, and we have to convince important players that it matters to them. Each area will be different. But here are some components that are likely to be common to almost all fields:

- Tools for creating and maintaining dictionaries
- Ways to extract information from raw sources (articles, papers, etc.) – that's why after the SePublica meeting we are "Jailbreaking the PDF".
- Getting authorities involved
- Tools to build and encourage communities
- Demonstrators and evangelists
- Stores for our semantic resources
- Working with funders

A small number of publishers do adopt this approach. I single out the International Union of Crystallography, which for many years has developed machine-understandable dictionaries for its discipline. Anyone publishing in their journals must submit in CIF format (Crystallographic Information Framework) which uses a name-value approach for data and a LaTeX-inspired approach for text. Papers are reviewed both by humans and machines, and the machines very frequently discover poor, bad, and sometimes fraudulent science. The final paper is automatically typeset from the (human-reviewed CIF). This is sufficiently compelling that a forward-looking publisher or society should surely be impressed.

So, apart from the political backdrop, why are semantics important?

- **They unlock the value of the stuff already being published**. There is a great deal in an a current PDF (article or thesis) that would be useful if it were semantic. Diagrams and tables are frustrating shadows of Plato's cave. Mathematical equations could be brought alive and computed in real-time by the reader ("plot that data, integrate the areas under the curves and compare with the equations"). Chemical structures can be extracted and their properties computed using Schroedinger's equation. Even using what we have today converted into semantic form would add billions.
- **They make information and knowledge available to a wider range of people**. If I read a paper with a term I don't know then semantic annotation may make it immediately understandable. What's rhinovirus? It's not a virus of rhinoceroses - it's the common cold. Semantic resolution makes it accessible to many more people.
- **They highlight errors and inconsistencies**. Ranging from spelling errors to bad or missing units to incorrect values to stuff which doesn't agree with previous knowledge. And machines can do much of this. We cannot have reproducible science until we have semantics.
- **They allow the literature to be computed**. Many of the semantics define objects (such as molecules or phylogenetic trees) which are recomputable. Does the use of newer methods give the same answer?
- **They allow the literature to be aggregated**. This is one of the most obvious benefits. If I want all phylogenetic trees, I need semantics – I don't want shoe-trees or B-trees or beech trees. And many of these concepts are not in Google's public face – we have to collect them.
- **They allow the material to be searched**. How many chemists use halogenated solvents? (The word halogen will not occur in the paper so Google can't find it). With semantics this is a relatively easy thing to do. Can you find second-order differential equations? Or Fourier series? Or triclinic crystals?
- **They allow the material to linked into more complex concepts**. By creating a data base of species, a database of geolocations and links between them we start to generate an index of biodiversity. What species have been reported when and where? This can be used for longitudinal analyses – is X increasing/decreasing with time? Where is Y now being reported for the first time?
- **They allow humans to link up**. If A is working on *Puffinus Puffinus* in the northern hemisphere and B is working on *Puffinus tenuirostris* in Port Fairy Victoria AU then a shared knowledge base will help to bring the humans together. That also happens between disciplines – microscopy can link with molecular biology with climate with chemistry.

None of this requires inferential logic. A hybrid mixture of terminologies, identifiers, data structures can be glued together into domain-aware systems. Semantics allow smart humans to develop communal resources to develop new ideas faster, smarter and better.

How do we make this happen? What I suggest may seem daunting but it's a smaller scale than the already successful Wikipedia. It is critical we act now**,** because Semantics/ContentMining is now seen as an opportunity by some publishers to "add value" by building walled gardens. If semantic enhancement is done by publishers then it will be very small, heavily controlled and expensive. So we must build something better, fully Open (i.e. no restrictions on re-use), and demonstrably valuable. It took one person to launch Open Street Map – and for many of us it's the gold standard of modern semantic maps – we can do the same for semantic publications.

We must create coherent communities. In the past this would be based on learned societies, but that will no longer work – we need a bottom-up approach. DBpedia is a beacon of how to create a world semantic resource – we must find ways of scaling this to disciplines it doesn't currently serve. It's conceivable that a mixture of the Wikimedia culture with public organizations (e.g. Galleries, Libraries, Museums, Archives) becomes the semantic core of scholarly publications.

Some semantic visions we should now be able to sell:

- **Give power to authors**. Authors are frustrated –many understand the need for annotations and are disenfranchised. Tools are becoming easier to deploy and we can create a semantic symbiosis for authors. For example a "species-checker" or "chemical checker" could be built into an authoring tool, so that the information is captured but the best person to understand it – the author.
- **Discover, aggregate and search** ("Google for science"). Search engines do not and will not support scholarly semantics. I cannot search Google for the details of numerical quantities, chemicals or species. It's relatively cheap and simple to do much of this – we indexed 500,000 reactions from US patents to a higher semantic quality than elsewhere.
- **Make the literature computable**. If we can compute parts of a paper we read, or aggregate many papers and map-reduce them, huge visions open up. For example we could search for all compounds in the literature which might sequester Carbon dioxide and compute their properties. This is a well-defined task and relatively straightforward to do.
- **Smart "invisible" capture of information.** If we interact with information (by creating it or reading it) then machines can also read and compute it. We would use semantic because they help us, but their results would be useful to the world. We use Bitbucket/Git because it helps us produce better programs, but a by-product is the archival for the whole world. Tools to help authors can also capture information seamlessly.

A critically important thing we can do now is to create a single-stop location for tools. Many new tools are being created and libraries such as Apache, Guava or UIMA contain much of what we need for simple conversion of raw material to semantic form. Key aspects are

- Common approach to authoring
- Crawling tools for articles, theses.
- Converters of PDF and Word to XML or XHTML
- Classifiers
- NLP tools and examples
- Diagram interpretation (e.g. extraction of data from graphs or phylogenetic trees)
- Logfile hackers (much output is "FORTRAN"-like and semi-structured). We can convert this to semantic form or annotate it automatically
- Semantic repositories
- Abbreviations and glossaries
- Dictionaries and dictionary builders

Scholarly publishing must change dramatically if only because the world is changing so fast. The present "mainstream" (traditionally closed-access) publishers cannot continue here as their model is to possess and control re-use of information, not to enhance it. The new semantic world will not only be formally Open but will think that way. Among the organizations (deliberately unnamed) that I expect to be responsive to the ideas expressed here are:

- Funders of science

- major Open publishers
- Funders of social change
- Open publication advocacy organizations
- (Europe)PMC
- Wikipedia
- GLAM
- Governments and NGOs

Where can a reader start? If they are in an organization, they can examine how semantic publication can enhance its business. If they are individuals they can build semantic tools and semantic resources.

And a remarkable example of the possibility was given in the post-SePublica hackathon ("Jailbreaking the PDF") – a collection of working tools and examples that show that current PDFs can often be transformed to semantic form.

Twenty-Five Shades of Greycite: Semantics for referencing and preservation

Phillip Lord, Lindsay Marshall

School of Computing Science, Newcastle University

Abstract. Semantic publishing can enable richer documents with clearer, computationally interpretable properties. For this vision to become reality, however, authors must benefit from this process, so that they are incentivised to add these semantics. Moreover, the publication process that generates final content must allow and enable this semantic content. Here we focus on author-led or "grey" literature, which uses a convenient and simple publication pipeline. We describe how we have used metadata in articles to enable richer referencing of these articles and how we have customised the addition of these semantics to articles. Finally, we describe how we use the same semantics to aid in digital preservation and non-repudiability of research articles.

1 Introduction

The academic publishing industry is changing rapidly, partly as a result of external changes such as the move to open access, and partly as a final recognition in the importance of the web. With change comes the opportunity to add more semantics to publications [1–3], to increase the computational component of papers, enabling publication to take its place in the linked data environment [4].

While within academia, third party publication — where knowledge is given to a third party to manage the publication process — is commonplace, outside in many technical disciplines we see direct publication, where the author publishes work that readers can then directly access. This form of publication is often called "grey literature" publication — a somewhat derogatory term — however, it has some significant advantages. It is rapid and places the author in control, allowing them to innovate in terms of presentation and content[5]. It operates without editorial control from third-party publishing which may help to overcome the publication bias found in many areas of scientific publishing. We have previously used a form of grey literature publishing to publish ontology tutorial material[6, 7]; this has resulted in the release of useful material which would otherwise probably not have been created, as many academics regard book chapters as having little purpose[8].

From the perspective of semantic publishing, it has an additional advantage; the process is often very simple, without additional human intervention between the author and the final published form. This simplicity means that semantics added by the author can pass through to the published version with relative

ease. In the process, it is also possible that semantics added by the author can aid in the authoring process, which we consider of key importance[9].

However, grey literature publishing lacks some of the formality of third-party academic publishing; for instance, several organisations provide centralised collection of bibliographic metadata; we have used this metadata, for instance, to enable accurate citation of academic literature through the use of primary identifiers. The lack of a centralised authority for author published literature, however, prevents this technique from being used for general URIs. This presents us with a simple research question: are there enough semantics on the extant web to provide clear bibliographic metadata for different web pages?

In this paper, we describe two new systems: greycite and kblog-metadata. The former, addresses the problem of bibliographic metadata, without resorting to a single central authority, extracting this metadata directly from URI endpoints. The latter provides more specialised support for generating appropriate metadata. We describe how these systems support our three steps doctrine [9], which suggests that semantic metadata must be of value to all participants in the publishing process including the authors. We also describe how these systems can impact on another major problem with author-led publishing: that of archiving and "link-rot".

2 References

Referencing is ubiquitous within scientific and academic literature, to the extent that it can be considered to be a defining feature. Academics reference previous work both as a utility to the reader, and as a mechanism for establishing provenance. However, reference insertion and formatting is complex to the point of humour[10]; with nearly 3000 citation formats in common use[11], reversing the process is even harder.

We have previously described the *kcite* tool which enables automatic generation of reference lists from a primary identifiers[9]: as described previously, it is often possible to hide these from the user behind tooling, so that they do not need to insert primary identifiers by hand[9]. This form of referencing also has advantages for human and machine consumption of the data; the primary identifier, which is also accessible to downstream analysis; moreover, because the reference is generated as a result of this identifier, when the author checks the reference, they are also effectively checking the identifier, which conventionally, the author must check manually at extra cost to their time.As an *ad hoc* measure, user feedback from our tool has now identified a number of primary identifiers (DOIs) with inaccurate metadata, and one systematic error in the presentation of these identifiers affecting many institutional repositories[12].

This, however, requires a source of metadata: currently, kcite supports (most) DOIs, arXiv and PubMed IDs directly, all of which allow metadata harvesting. Following the development of kcite, our request resulted in both CrossRef and DataCite – the two most significant DOI registration agencies for academia – providing metadata in the form that kcite consumes (Citeproc JSON). For gen-

eral URIs, unfortunately, there is no centralised authority which can provide this metadata.

2.1 Technical Glossary

Here, we provide a short technical glossary of the tools described, also shown in Figure 1, as an aid to understanding.

kcite: A wordpress plugin that generates a reference list for an article from primary identifiers. Uses a variety of services, including greycite, to resolve identifiers to bibliographic metadata.

kblog-metadata: A wordpress plugin that provides flexible presentation of bibliographic metadata, both computationally and visibly through on-screen widgets.

greycite: A server which returns bibliographic metadata for any URI, extracted from that URI for the article resolved by that URI.

Citeproc JSON: A bibliographic format defined by the Citeproc-js tool.

BibTeX: a format defined by the BibTeX tool.

3 The Greycite System

We initially considered the possibility that kcite could harvest its own metadata directly. It would have been possible, for instance, for a kcite installation on one site to return metadata to another, through a REST call, or as embedded metadata. However, this would have required users to know in advance which URIs were so enabled, and would have worked with few websites.

To avoid this limitation, we wished to use extant semantics already on the web; the complexity of this task argued against integration with kcite which is an end-user tool. Additionally, as a server greycite would usable to more than one client; in fact, this has proven to be the case, with a third-party tool, knitcitations which supports dynamic citations in a literate programming environment for R[13].

Greycite provides bibliographic metadata in a variety of formats on request about an arbitrary URI; an architectural overview is shown in Figure 1. It uses a simple REST API to do this, and returns either Citeproc-JS JSON (directly used by kcite)[14], BibTeX (used by knitcitations, and the kblog-metadata tool described here). We have additional support for other formats, including RDF (encoding Dublin Core), RIS, and Wikipedia "cite" markup. It store the results of metadata extractions, initially for reasons of efficiency, although this is also valuable for ephemeral sources of metadata(see Section 4).

Greycite extracts a number of sources of metadata, and uses a scoring scheme and a set of heuristics to choose between them; we describe these next.

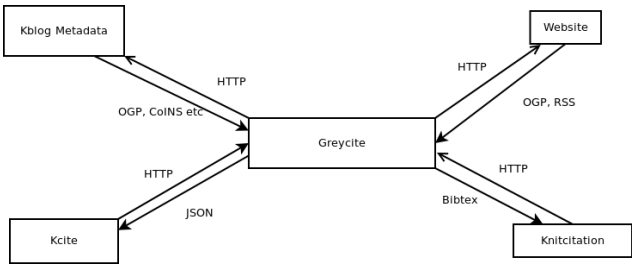

Fig. 1. Client server interaction between Greycite and clients

4 On how the Web describes itself

To enable referencing, we need five key pieces of bibliographic metadata, namely:

- Author(s) (A)
- Title (T)
- Date of publication (D)
- "Container" – equivalent to journal, conference or website. (C)
- Canonical Identifier (I)

These are the minimal pieces of metadata used by most referencing styles, and following standard publication practices. We have now investigated many sources of web-delivered metadata. These have been discovered in a number of ways: some were designed for this purpose. Others, where discovered by inspection of academic websites; some were discovered entirely by chance (where an authors name was visible on a web page, but not extractable, we search for all instances of that name, looking for structure). We prioritised "interesting" websites, for our definition of interesting.

A complete list of all the mechanisms greycite uses for metadata extraction is shown in Table 1. By itself HTML provides very few of these five pieces of metadata; only the title is extractable; even here, for most browsers, the title is displayed publicly, in the browser title bar. As a result, many sites include the name of the site, often "—" delimited in the title of each page, which makes this a relatively messy form of data.

We also investigated the use of CoINS metadata; this standard is used by a number of academic websites, and can be consumed by a some bibliographic tools[1]. It is an imperfect tool. The standard is rather confusing to read, the main website describes it as using a NISO 1.0 Context Object, the link to the specification for which is broken. Different implementations tend to produce different variations of the same metadata. More over, CoINS metadata does not necessarily describe the article being posted; for example, `http://researchblogging.org` uses CoINS to describe a secondary article being reviewed. It has a significant

[1] `http://ocoins.info`

advantage, however, over most metadata specifications which is that it is embeddable in the *body* of a web page; for hosted websites, authors often do not control the headers and cannot add elements to it.

The guidelines for inclusion to Google Scholar are somewhat clearer, and easier to implement, although even here there are common causes for confusion (`citation_author` vs `citation_authors`). This form of metadata is relatively common on many journal websites, but is not, in our experience, wide-spread outside academia. More common, is Open Graph Protocol, or OGP[2]; this is a form of RDFa developed as part of the Facebook platform. It is found on a large number of websites including many common blog platforms, as well as various news outlets, such as BBC News, which are otherwise hard to cite. The author list is often not represented in OGP[3]; while OGP has the ability to do this, authorial metadata needs to be gathered from a secondary URI, linked from the main content; this is more complex to implement, which may explain why it is commonly missing.

Another source of authorial metadata is RSS/Atom feeds. Many common platforms include a `dc:creator` tag and this is often the only easily extractable form of metadata. We do find that generic (`admin`, `blog admin`) or personal but informal (`Phil`, `phillord`) user names are fairly common; this is the default behaviour for many content management systems, and appears to be a conscious choice for many multi-user sites. Greycite filters some of the more common ones and does not consider them as valid metadata. We also provide heuristics where articles are missing; for instance, if all articles in an RSS feed have the same author and container title, we infer this information for missing articles.

Another commonly missing piece of metadata is date; while it can be found in RSS/Atom feeds, these are not always present and are *ephemeral*. In contrast to author or container information, publication date cannot be infered where articles are missing from metadata on other articles. We apply a heuristic here in acknowledgement of the fact that many blogs use a date format for their URI permalinks. In fact of the URIs in greycite, we can mine date metadata from some 33% of them; while this is not a representative sample, it does show that heuristics can be surprisingly effective.

Unfortunately, many scientific papers are published in PDF; while we do attempt to extract metadata from these, greycite is currently not very effective, so most PDFs appear to contain no extractable metadata; we are investigating more PDF parsers to attempt to address this problem. In some cases, we have provided heuristics which work around this difficulty: greycite will provide metadata for PDFs hosted by CEUR-WS; however, we achieve this by mining metadata from the HTML files which link to the PDF.

A significant number of websites do not provide any specific metadata that we were able to discern; interesting and surprising cases include most of the W3C standards, websites for both the International and Extended Semantic Web Conferences, and the ORCID webpages. We have a significant number

[2] `http://ogp.me`
[3] Including on the OGP website!

Source	Type	Notes
Atom	TDCAI	Inferences where article is not present
CoINS	TCDA	Blocked where identifier does not match location
CEUR-WS	TCDA	Uses span tags in index files
Dublin Core	TCDA	Both `dc:` and `dc.` recognised
Eprints	TCDA	
EXIF	TDA	In Progress
FOAF	N/A	In Progress
GIF	N/A	In Progress
Google Scholar	TCDA	Both `citation_author` and `citation_authors`. Bepress prefix with `bepress`
HTML	T	The "title" tag
Link	N/A	In Progress
Meta	TCDA	Common uses recognised
OGP	TCDAI	Some syntactic variants
OpenLibrary	TCDI	Preliminary
ORCID	TCA	Screen Scraping
PDF	TA	Often fails!
Prism	CD	
RSS	TCDAI	See Atom
Schema	TD	In Progress
Scholarly HTML	N/A	In Progress. Never seen in the wild
ScienceDirect	TCD	Screen Scraping
Twitter	TCAI	"Author" is normally a hashtag
URI	D	Heuristic based on link structure
W3C	TCDAI	Screen scraping specific for W3C specifications
WorldCat	TDA	Screen scraping

Table 1. Twenty-Five Sources of Metadata: type indicates the metadata extractable (**T**ype, **D**ate,**C**ontainer,**A**uthor,**I**dentifier).In progress indicates that we believe more metadata is present. Screen Scraping means heuristics based on HTML structure.

of special purpose extraction plugins; for instance, from a desire to reference W3C specifications, we have created a single site plugin which uses a highly *ad hoc* screen-scraping technique. Taken together, of the 4000 URI that have been submitted, Greycite can extract the main four pieces of metadata (TCDA) from 62% of URIs.

5 On how the web could describe itself

While Greycite can extract metadata from many different sources, it does require some support from the content. Unfortunately, for many content management systems whether this metadata is available or not is dependent on the local setup; for instance, with WordPress, the presence or absence of many sources of metadata is theme dependent; the exception to this is data from the RSS/Atom feeds although even here, the feeds themselves can be disabled at the theme level[4], or through author choice[5].

We have therefore created a plugin for WordPress to address this need; while the solution is, of course, specific to WordPress, the use cases that we address are considerably more general. This plugin, *kblog-metadata*, currently adds metadata in three formats: Google Scholar, OGP and CoINS. The latter is used by and has been tested with Zotero and similar bibliographic tools, which is the main reason for its inclusion. Facebook provide an explicit tool for testing OGP, while Google Scholar do not. By default, kblog-metadata requires no configuration and uses knowledge directly from the WordPress container, which provides suitable values for the five pieces of metadata we require (see Section 4).

While the author can check that their metadata is appearing correctly, through the use of greycite, this requires them to use a secondary website. Alternatively, they can link to their article using kcite, which will then generate a reference on the basis of the metadata; however, this requires creating new content, to check old. Following our three steps doctrine, we wished to make the metadata more useful for the authors. We have, therefore, added "Widget" support, which displays citation information for each page (or a website as a whole) using the same metadata resolution techniques; this display both eases the task of checking the metadata, as well as incentivising the author to do so. The widget also provides a BibTeX download of the citation. As well as being useful for authors and readers, this has an additional utility: the BibTeX actually comes from greycite, on the basis of its metadata extraction. When anything (including robots) access this BibTeX, Greycite is invoked, and hence becomes aware of the new article.

Although for simple use, the default WordPress data suffices, there are several uses cases where it does not. Therefore, kblog-metadata provides authors with the ability to set the metadata independently on an individual post basis. This fulfils a number of use cases. The most common of these is for multiple-author posts; WordPress multiple author support is built around editing rights, rather

[4] This would generally be considered to be a broken theme
[5] This would generally be considered to be a broken author

than authorship. Hence all authors must have WordPress logins which they otherwise may neither want or need. Kblog-metadata allows setting authorship lists independently of login rights. Secondly, authors may also wish to provide an alternative container title. Combined, these two facilities enable WordPress to operate as an "preprints" server. For example, `http://www.russet.org.uk/blog/2054` resolves to the full text of our paper[9], which uses both facilities so that the citation appears with three authors, and "Sepublica 2012" as the container title.

Since, kblog-metadata was released, WordPress also supports "Guest authors" through the co-authors-plus plugin – which likewise dissociates login rights from authorship; this provides a much nicer graphical environment for defining co-authors than kblog-metadata, but comes with an overhead that authors must be created individually. Kblog-metadata will use metadata from this plugin if it is installed.

Finally, we have added support for the use of shortcodes to define authorship. This is very useful when content is being generated outside of the WordPress environment; for example, on `http://bio-ontologies.knowledgeblog.org`, most of the content is generated from Word documents. During publication, we markup the author names with shortcodes — `[author]Phillip Lord[/author]`; this markup passes unmolested through Word's HTML conversion and is then interpreted by WordPress. This prevents cut-and-paste errors that would occur if authors had to be added manually — a significant issue for science where most articles have many authors. This website also modifies the container title to distinguish between different years.

6 Identifying by Proxy

One significant issue with kcite as a referencing engine is the requirement for a primary identifier for every item[6]. Most scientific literature, and any article posted on the web is likely to have an identifier that kcite can use. However, this causes problems for two specific types of resource. First, many smaller conferences and workshops do not publish their literature in a web capable form; in many cases papers on the web are available as PDF or Postscript only. And even when web hosted, sites may not add bibliographic metadata. Kblog-metadata provides a partial solution to these problems: authors can host their articles, and alter the metadata accordingly as described in Section 5. However, this fails for work by other authors, whose work cannot be posted without permission. A similar problem exists for books; while these generally do have a standard identifier (ISBN) we have not been able to find a publicly available mechanism to automate the transformation from ISCN to structured bibliographic metadata.

Greycite provides a mechanism to address this difficulty. There are a number of catalogues available for both scientific literature and books; these often

[6] kcite does allow addition of all citation metadata within an inline shortcode, although this is intended as a fallback

have a primary URI which can be used as a reference identifier. Greycite currently supports several sites of this form: WorldCat (`http://worldcat.org`) provides URIs for books (as well as other forms of media such as CDs and DVDs), Mendeley which references journal articles and OpenLibrary (`http://www.openlibrary.org`) which also provides URIs for books. In these cases, references will appear correctly when used in Kcite, showing the source of metadata which could, in principle, be used to track the original resource.

7 Metadata for Preservation

One recurrent issue with author-led publishing is the difficulties associated with digital preservation; custom and practice means that it is considerably harder for author-led publications to ensure that work is preserved than third-party publications; systems such as CLOCKKS or LOCKKS are often just not accessible to smaller-scale author-led publication.

To address this need, we have integrated greycite with public archiving efforts, such as the Internet Archive, the UK Web Archive and WebCite. As well as scanning URIs for metadata, greycite periodically checks these archive sites, to see if they are available as archives. We use this metadata in a number of ways.

First, archive sites are available directly from Greycite through a REST API. Kblog-metadata provides an "archived" widget where it publicly displays this information; this provides a third-party stamp that the article has actually been available from the time stated, as well as an *ad hoc* enforcement of non-repudiability. If an author changes their own content, the differences with the archived sites will be clear.

If a site disappears, then these links to the archives will also disappear. Kblog-metadata also allows readers to download BibTeX files for any (or all) articles; this metadata comes directly from Greycite and includes links to all known public archives. Anyone citing an article using this file will therefore have a reference to archival versions.

Of the services we currently check for archival versions, currently only WebCite offers on-demand archiving[7]. Greycite currently submits any archive with four (TCDA) piece of metadata to WebCite for archiving. Additionally, greycite itself stores historical metadata for indeterminate amounts of time, and therefore constitutes a metadata archive.

8 Tracking movement around the web

In addition to the four pieces of bibliographic metadata, greycite collects one other key piece of knowledge; a canonical URI. Currently, this knowledge is not represented in many of the formats we harvest. While, CoINS does provide a field which can be used for this purpose, in practice it is not that useful: CoINS is used

[7] WebCite is asking for funding on the web, which is an unfortunate sign

to embed bibliographic metadata into the web, but the CoINS may not relate to the article in which it is embedded. Open Graph Protocol data also returns an explicit identifier; in this case, this is about the article in question. This means Greycite can store a canonical URI for a particular article, regardless of the URI used to access the article. Again, and perhaps unexpectedly, RSS/Atom feeds are extremely useful; these carry a link and explicitly state whether it is a permalink (i.e. canonical) or not.

The presence of a canonical URI makes it possible to track content as it moves. For instance, it is relatively common for blogs to change their permalink structure; with WordPress, for instance, existing links are maintained through the use of a 301 Redirect response. Greycite could recognise this situation and use the Redirect location as the canonical link; unfortunately HTTP redirects are used for many different purposes, including load balancing. Instead, greycite recognises that the URI used to fetch a request and the stated canonical URI are different and records this fact. For example, greycite records that `http://www.russet.org.uk/blog/2012/02/kcite-spreads-its-wings/` changed from being canonical to not sometime between April 2012 and Jan 2013 (actually this happened in June 2012).

Currently, greycite returns the canonical URI with requests for both BibTeX or JSON data; authors of referring documents will therefore will have an recent link. Although, currently not implemented, we plan to add more explicit support for this to our kcite client, so that it will display canonical URIs; again, this supports digital preservation. Articles which refer to URIs which have ceased to be canonical, would be able to display both the URI to which the author original referred and the correct canonical reference.

The ability to track articles as they move also opens up a second possibility. Currently, one main stated advantage of systems like DOIs is the ability to change the location of a record without necessitating a change in identifier. A similar system is also available in the form of PURLs (persistent URLs)[8]. Greycite allocates PURLs for all URIs for which it can extract all the required piece of metadata. Currently, these redirect to the last known canonical URI for a given URI; in effect, this means that PURLs will track URIs for any website that maintains its redirects and metadata for sufficient time for Greycite to discover this.

9 Discussion

As we have previously stated[9], our belief is that semantic metadata, if it is to be useful at all, must be useful to all the key players in the publication process; critically, this includes the author. The tools that we have described here obey this doctrine; we seek to aid and reward the authors who use good metadata.

Kcite already follows this principle: if links are inaccurate, then the reference will not format correctly (or at all). As well as errors made during authoring, we

[8] `http://purl.oclc.org/`

(PWL) have found non-functioning DOIs, as well as one systematic error in DOI presentation which has resulted in a change to CrossRef display guidelines[12]. In addition, correct formatting of the references depends on the metadata being correct. Again, here, we have found DOIs with inaccurate metadata. With the addition of greycite, this functionality has been extended to any URI. Authors are very likely to cite themselves. If they do so, they are now dependent on their own metadata; if the metadata is wrong, then references will be. This provides an incentive for authors to correct metadata for their own purposes, simultaneously making everyone's life better[9]. As well as correcting our own websites, use of greycite has discovered inaccurate metadata in commercial publishing websites.

Greycite is currently unique so suffers from some of the limitations of centralisation; however, effectively, it is just a cache. The metadata that it provides is sourced from the distributed resources that are referenced; it can support multiple installations trivially. Except in the case of ephemeral metadata, none of these would be privileged. The current implementation of greycite also provides an initial answer to our question, is there enough bibliographic metadata on the web to enable citation: a qualified yes. Through the use of existing metadata schemes and some heuristics, we can discover this metadata for many websites. An early analysis suggests that greycite can provide the four key pieces of metadata for around 1% of the web, which consitutes 100s of millions of URIs; the percentage for "interesting" websites is much higher, at over 60%. We currently also lack any statistical analysis on how correct this metadata is; by inspection, the level of correctness within the ∼4000 URIs submitted from 254 independent IP addresses is high, but this result is biased as we have corrected errors iteratively. For random URIs, we lack a gold standard, and most are not in English making inspection hard.

While using metadata to generate references is useful, it is one-step removed. The author is not supported in discovering that their metadata is inaccurate until sometime after it has been published. Kblog-metadata now improves on this process and makes it more immediate; by visualising the metadata on publication, authors can check that it is correct. Likewise, the same metadata is used to generate a BibTeX file which they can use. As an open source tool, it is hard to know how many installations kblog-metadata currently has, although download statistics would suggest 30 or 40, including one journal.

Set against this desire to improve the quality of metadata on the web, greycite has taken a pragmatic approach to the metadata standards it uses. It currently supports many of the different ways of marking up bibliographic metadata. More over, it uses many heuristics, to cope with metadata which is unclean or just broken. This works against the notion of encouraging authors to improve their metadata; however, increasing the utility of the API makes this a compromise well worth making.

We are also addressing the issue of digital preservation; we achieve this in two ways. First, we leverage existing web archives, deep linking through to them where content has already been archived. To achieve this in a simple manner

[9] slightly

requires no semantics at all, beyond the URI for a given resource. However, resources may be present at more than one URI, or may change their canonical URI over time. Greycite is now making preliminary use of this metadata to track articles as they move; the current location can be retrieved by a client, or alternatively greycite provides PURLs which will work with any client.

As with our previous work, the level of semantics provided or used by these publication tools is not high; however, by using existing metadata standards, greycite can provide metadata for 100s of millions of URIs including many from websites which are unlikely to care about academic referencing. We have focused on adding value for authors, both when referencing or displaying citations on an article. By adding value for the authors, we help to ensure that they will add value to the metadata. While this approach adds very small amounts of metadata for an individual article, the aggregate total of metadata over all articles is, potentially, vast.

References

1. Shadbolt, N., Hall, W., Berners-Lee, T.: The semantic web revisited. Intelligent Systems, IEEE **21**(3) (2006) 96–101 http://eprints.soton.ac.uk/262614/.
2. Shotton, D.: Semantic publishing: the coming revolution in scientific journal publishing. Learned Publishing **22**(2) (2009) 85–94 http://delos.zoo.ox.ac.uk/pub/2009/publications/Shotton_Semantic_publishing_evaluation.pdf.
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 http://dx.doi.org/10.1371/journal.pcbi.1000361.
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 http://eprints.soton.ac.uk/271285/.
5. Zhu, Y., Procter, R.: Use of blogs, Twitter and Facebook by PhD Students for Scholarly Communication: A UK study. In: China New Media Communication Association Annual Conference. (2012) https://www.escholar.manchester.ac.uk/item/?pid=uk-ac-man-scw:187789.
6. Lord, P.: Ontogenesis. http://ontogenesis.knowledgeblog.org/ (2012)
7. Lord, P., Cockell, S., Swan, D.C., Stevens, R.: Ontogenesis knowledgeblog: Lightweight semantic publishing. http://www.russet.org.uk/blog/1920 (2011)
8. Bishop, D.: How to bury your academic writing. http://deevybee.blogspot.com/2012/08/how-to-bury-your-academic-writing.html (2012)
9. Lord, P., Cockell, S., Stevens, R.: Three steps to heaven. http://www.russet.org.uk/blog/2054 (2012)
10. Psyphago, .: Scientists receive 12.6 million dollar grant to format references correctly. http://collectivelyunconscious.wordpress.com/2013/02/27/scientists-receive-12-6-million-dollar-grant-to-format-references-correctly/ (2013)
11. Zelle, R.: Styles. http://citationstyles.org/styles/
12. Lord, P.: The evil a space can do. http://www.russet.org.uk/blog/2340 (2013)
13. Boettiger, C.: Semantic citations for the notebook and knitr. http://www.carlboettiger.info/2013/02/22/semantic-citations-for-the-notebook-and-knitr.html (2013)
14. Bennet, F.: The citeproc-js processor (2009) https://bitbucket.org/fbennett/citeproc-js.

Systematic Reviews as an interface to the web of (trial) data: Using PICO as an ontology for knowledge synthesis in evidence-based healthcare research

Chris Mavergames[1], Silver Oliver[2], Lorne Becker[3]

[1]The Cochrane Collaboration, Director of Web Development, Freiburg, Germany
mavergames@web.cochrane.org
[2]Ontoba, Information Architect, London, United Kingdom
silver.oliver@ontoba.com
[3]Cochrane Innovations, Director, Tampa, USA
lornebecker@gmail.com

Abstract. Linked data and semantic technologies offer flexible and powerful solutions for connecting and synthesizing knowledge in the domain of healthcare research, in particular the area of evidence-based medicine (EBM). Systematic reviews of healthcare interventions, the primary methodological approach for evidence synthesis in EBM, involve a rigorous and time-consuming process of collecting and analyzing data from all the studies conducted around a particular clinical question. The number of primary studies reported each year is rising exponentially, and the process of producing systematic reviews that synthesize data from these studies is labor-intensive and keeping these reviews up to date a huge challenge. Currently, the reviews are primarily published in PDF format and much of the value is locked away from programmatic access. This position paper discusses the potential in using linked data technologies to improve discovery of knowledge in systematic reviews by using the PICO (Population, Intervention, Comparison, Outcome) framework as an ontology to aid in knowledge synthesis.

Keywords: linked data, systematic reviews, randomized controlled trials, Cochrane Collaboration, Cochrane Library, semantic web, Drupal, RDF, SPARQL, OWL, ontologies

1 Introduction

The evidence-based approach to health care decision-making calls for clinicians to find and apply the best and most up-to date results of medical research in their clinical decision making. This is no easy task, even if the search for evidence is restricted to the highest quality evidence (such as that provided by randomized controlled trials or RCTs). Thousands [1] of RCTs are completed each month and most result in several papers or reports which are often published in different journals and may not reference each another. Added to this, different RCTs addressing the same clinical question may have widely differing findings.

A relatively new type of approach, the Systematic Review, has been developed to address this problem. Systematic Reviews of healthcare interventions use rigorous methods to identify all of the studies that have investigated a particular clinically-relevant question, appraise their methods and combine their results to present a synthesis of the evidence for the question of interest. The PICO format is used to frame the clinical question as to the Population (patient with problem), the Intervention (drug or other) being given, the Comparison of that intervention with another intervention, no treatment or a placebo, and in relation to which Outcome(s) (symptoms relieved, etc.). Annotation of data and content by PICO forms a powerful framework for navigating and synthesizing the evidence and can help to create an interface onto the web of trial data available for analysis.

1.1 The Cochrane Collaboration and systematic reviews

The Cochrane Collaboration (http://www.cochrane.org), an international, non-profit research organization has developed much of the methodology underpinning the systematic review approach and publishes The Cochrane Database of Systematic Reviews (CDSR), which now contains more than 5,200 of these innovative articles. Although each Cochrane Review can be downloaded as a PDF, the CDSR was originally conceptualized as a database and has never been published as a paper journal. At its inception, almost 20 years ago, the CDSR was distributed on floppy disks, transferring to an online format once the World Wide Web came of age. Cochrane Reviews are continually updated and previous versions of the Reviews are available for download. Thus, one can track the "status" of the answer to a particular clinically-relevant question over time.

Making Cochrane reviews as accessible as possible is at the core of Cochrane's remit. Cochrane aims to help

people make well-informed decisions about health care by preparing, maintaining and promoting the accessibility of systematic reviews of the effects of healthcare interventions. [2] While Cochrane Reviews are internationally recognized for their rigor, validity and unbiased synthesis, they are long and complex documents that can be intimidating for naïve or even for somewhat experienced users. [2] There are Reviews that span 800-pages and ones with more than 750 analyses in the form of forest plots (http://en.wikipedia.org/wiki/Forest_plot).

Recently, Cochrane began a linked data project to investigate how semantic technologies could assist in helping users navigate the evidence in Cochrane Reviews, while also assisting in the process of producing them. The project hoped to build on the work already done in this space around creating ontologies for clinical research (OCRe - http://bioportal.bioontology.org/ontologies/1076) as well as work done to convert clinical trial reports into RDF (http://linkedct.org). The aim of the linked data prototype was to improve access to Cochrane Reviews through a demonstrator website orientated around the user stories and questions that user research showed clinicians and other end-users of Cochrane evidence are seeking to answer. Emphasis was placed on new access points for search and navigation and PICO emerged as the primary framework for tagging the content.

1.2 PICO

The PICO framework has long been held as a key mechanism for information retrieval in evidence health care. [3] PICO stands for Population (patients with a condition), Intervention, Comparison and Outcome(s). For example, is drug A (Intervention) effective for the relief of B condition in C Population in Comparison with X drug (or placebo) for Y Outcome(s) (symptoms relieved, etc.). Cochrane reviews use the PICO framework extensively at each stage of formulating a review: question, searching, screening, analysis and publication. PICO was identified as a key method of providing access points and facets for browsing in the Cochrane linked data demonstrator.

Research into using PICO for information retrieval [3,4] has affirmed its usefulness and explored options for identifying these elements in corpora of documents via manual annotation and natural language processing techniques. The following aspects of the PICO framework from this research [3,4] were drawn upon in the design of the demonstrator:

- Weakness of expressivity in terms of relating PICO elements
- Overloading of P (condition, aspects of the population)
- Overriding benefits of P and I (as opposed to C and O) to retrieval.
- Works better for therapeutic questions

Thus, the prototype looked to focus on the following:

- Creating a Cochrane Review ontology that also covered the studies that are included in the reviews.
- Modeling PICO within this ontology of reviews and studies by breaking the elements of PICO into more specific classes and their relationships.
- Drawing on linked data sets for populating the PICO elements with instances from controlled vocabularies such as SNOMED CT.

2 The ontology

A prototype ontology (Fig. 1) that partially models Cochrane Reviews, the studies that are included in them, and the PICO framework was created. The purpose of this ontology was to fulfil the specific set of user stories drawn from user research into the kinds of questions people were trying to answer using Cochrane systematic reviews and to drive improvements to navigation and discoverability of content within the documents. The prototype ontology was developed from a “product” (published Cochrane Review) perspective and thus the design was not driven by the underlying methodology of doing systematic reviews.

While care was taken to ensure that it is methodologically sound in its design, its purpose was to demonstrate the advantages of using a linked data approach to describing, storing and managing our content and processes. The next iteration of the ontology will be methodology driven and will likely result in the creation of a suite of ontologies to power evidence synthesis.

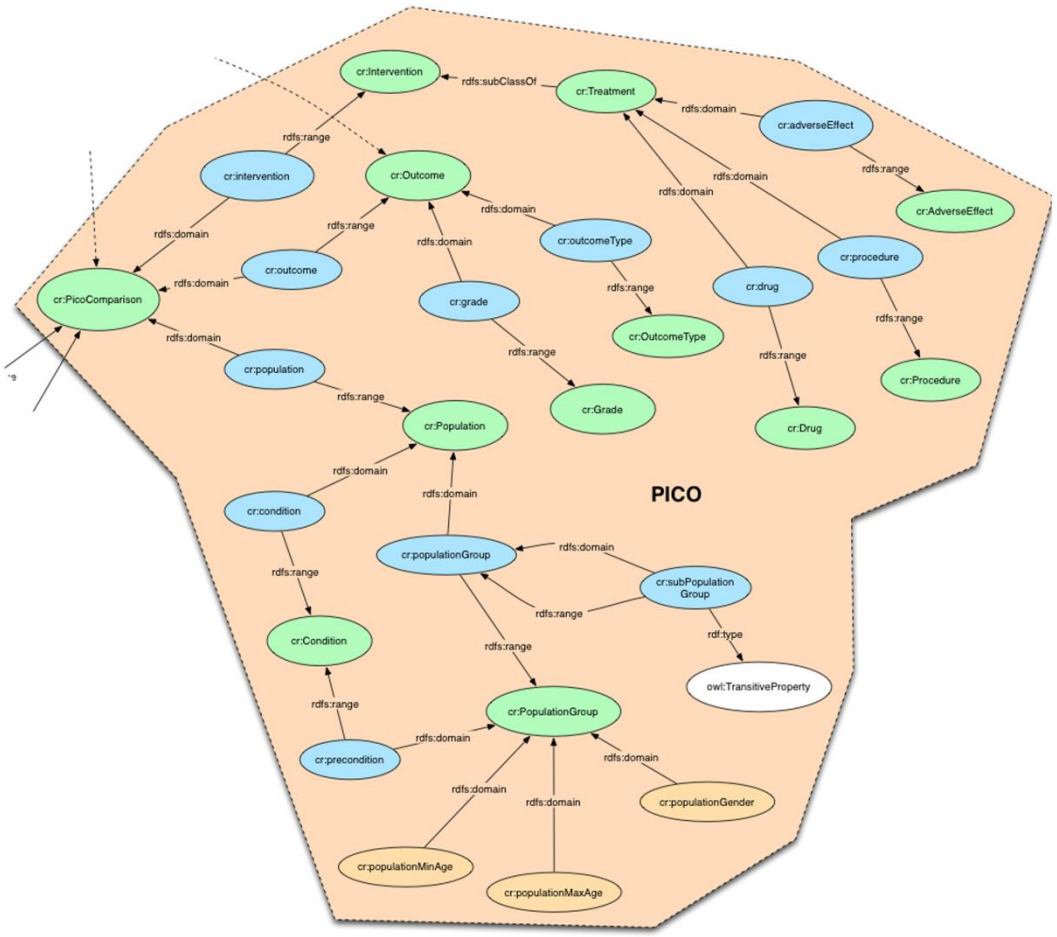

Fig 1. PICO portion of the Cochrane Review prototype ontology

3 The demonstrator

3.1 Information architecture

For the demonstrator [5], we focused on a small corpus of Cochrane Reviews on Asthma that addressed PICO questions where the interventions assessed were drugs. The information architecture of the site aimed to create:

- **Page (URI) per condition, intervention, study, review.** The demonstrator exposed the PICO model as part of the information architecture of the site. For core classes of concepts, each had a specific URI and links to related things in the graph. This encouraged browsing through rich user journeys. This was also considered beneficial from a search perspective exposing the key classes to be landing pages for external search.

- **Searching Reviews by drug name.** Currently, there is no cross-indexing against variant names of drugs in Cochrane Reviews. We have linked to Drugbank (http://www.drugbank.ca/) which includes most of the variants of drug names including the different brand names and generic names used in different countries. We created a "semantic search" that allows users to type any name for an asthma medication and find the relevant Cochrane Reviews. This functionality would greatly improve the discoverability of Cochrane content in The Cochrane Library (http://www.thecochranelibrary.com) as, for example, if you search for "Prozac" you get zero results, but if you search for "fluoxetine" you get 30 results.

- **Displaying selected portions of reviews.** Clicking on any title on the "List of Reviews" page in the demonstrator takes you to a custom view of that review that could eventually be customized based on

the query or route of browsing by PICO, once the ontology is fully populated with instances for tagging these elements.

- **Linking out to selected content.** In addition to linking to Drugbank as noted above, we have linked to SIDER (http://sideeffects.embl.de/), a linked data set that includes information on side effects from FDA label information, as well as linking to PubMed records via PubMed identifiers supplied by the Cochrane Register of Studies (CRS), Cochrane's "studified" register of RCTs published in CENTRAL (Cochrane Central Register of Controlled Trials), the largest collection of RCTs in the world.

3.2 Building the demonstrator site

The following steps were taken in building the demonstrator site [5]:

- **Development of user stories.** The project began with a relatively detailed analysis of key user stories. So all features built upon a sound understanding of business and user needs.
- **Development of a model.** A RDF model was developed to support the prototype. Here the majority of the focus was the interplay between the PICO elements and associating them with the right section of the review.
- **Extract Transform Load (ETL).** Scripts (Java) were created for transforming the XML review documents into RDF. Taking the XML and structure and minting URIs where necessary.
- **Annotation of reviews.** The reviews were hand annotated against the URIs populating PICO classes for I and C for Asthma drugs and interventions. This was done in a spreadsheet and converted to RDF using Open Refine with the Deri RDF plug-in. The annotations were at the question level of the review.
- **Build the "views".** The open-source content management system Drupal was used as the prototyping engine for the purpose of creating a demonstrator site. Drupal "plays nicely" with the semantic web stack including an RDFx module with a module called SPARQL Views which allows for SPARQL queries to be constructed within the core Drupal Views system. OWLiM was used as the triple store software and Drupal connected to this repository via SPARQL Views and queries and results were generated and rendered within the Drupal website. This allowed us to quickly create a working website that can be quickly styled and made functional using Drupal's built-in theming and templating system.

4 Discussion

4.1 PICO as an ontology

Though the work was only a proof of concept, and the development of the model is still evolving, the work to date has raised a number of significant benefits of capturing PICO as a rich RDF model.

- RDF lends itself well to rapid prototyping: the flexibility of an RDF data store and the ability to play with a variety of linked data sets to see how they enrich features. For example including RDF data from SIDER without having to do any additional modelling.

- Inference: using transitive taxonomical relationships in the drug class hierarchy to address issues of intervention queries at different hierarchy depths. For example Salmeterol having a child relationship with respect to Long-Acting Beta 2 Agonists. The relationship that associates the role of a drug in an intervention was transitive-over the hierarchy and thus this reasoning insured any query for Long-Acting Beta 2 Agonists would include PICO's referencing specific drugs such as Salmeterol.

- Rich queries: SPARQL queries taking advantage of the rich graph of PICO data showed great potential. Queries that would ask for a drug (including brand named drugs via Drugbank at any depth of the drug hierarchy) that has been compared to any other intervention for a particular condition go significantly further than equivalent information retrieval approaches that use PICO as a source of identifiers for classifying a document.

- Ability to map out to standard models and upper-level ontologies: RDF lends itself well to providing a mechanism to express the specific needs of the product in hand and map out to existing models for interoperability moving forward.

For the purposes of the prototype, PICO was only associated with the question level of the review and only I and C were used for annotation thus far. In addition, we identified at least three key places for the annotation of PICO with relation to Cochrane systematic reviews that should be explored and implemented going forward:

- Study level: PICOs being reported in the trial reports
- Question level: The PICO question posed by the systematic review
- Findings level: The results from analyses and meta-analyses for a particular PICO question in the review

And, further development of a stand-alone, PICO ontology will model, in detail, the facets of the various components of the PICO framework, in order to capture the richness and granularity present, especially with regards to P (population) and O (outcomes).

Each of these levels of metadata would provide increasingly more value to end-users in terms of discoverability, navigation and traversing the graph of available information and analyses. For example, users would be able to restrict searches for reviews to those that actually find results for a given question as opposed to simply where this question has been asked but without result. There is clearly a cost-benefit analysis to be done here, but machine-supported annotation has been explored for the labour intensive job of annotating at the study level, where studies for a given review might extend to hundreds.

5 Conclusion

Our experience with the Cochrane linked data project to date has convinced us that it has potential to become an "enabling technology" for the Collaboration that could allow us to do more with our data in terms of synthesizing evidence, but also to enable better discoverability and presentation. However, there are a number of issues that should be explored as we decide on how best to leverage linked data and semantic technologies both within the Cochrane technology stack. Also, there is possibility of moving toward an "operating system" for evidence-based healthcare research by leveraging semantic technologies.

The systematic review is an excellent example of the article as an interface to the underlying web of trial data. The use of PICO as an upper-level ontology for annotating studies and systematic reviews holds great promise in moving toward "living systematic reviews", whereby the notion of the systematic review as a document slowly fades in favour of "high quality online evidence summaries that are dynamically updated as new evidence becomes available". [6]

References

1. Bastian H, Glasziou P, Chalmers I (2010) Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? PLoS Med 7(9): e1000326. http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.1000326
2. Mavergames, C. Sustainability and Cochrane Reviews: How Technology can Help Plenary talk by Chris Mavergames from UK Contributors' Meeting in Loughborough, March 2012 http://www.slideshare.net/mavergames/sustainability-and-cochrane-reviews-how-technology-can-help-12207716.
3. Huang, X., Lin, J., Demner-Fushman, D. Evaluation of PICO as a Knowledge Representation for Clinical Questions. AMIA Annu Symp Proc. 2006; 2006: 359–363. PMCID: PMC1839740
4. Boudin, F., Nie, J., Dawes, M. Clinical information retrieval using document and PICO structure In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (2010), pp. 822-830.
5. Cochrane Linked Data Demonstrator, http://demonstrator.dev.cochrane.org
6. Elliott, J., Mavergames, C., Becker, L., Meerpohl, J., Thomas, J., Gruen, R., Tovey, D. Achieving high quality and efficient systematic review through technological innovation. BMJ Rapid Response. January 2013. http://www.bmj.com/content/346/bmj.f139/rr/625503.

Towards Linked Research Data: An Institutional Approach

Cord Wiljes[1], Najko Jahn[3], Florian Lier[2], Thilo Paul-Stueve[2], Johanna Vompras[3], Christian Pietsch[3], and Philipp Cimiano[1]

[1] AG Semantic Computing, CITEC, Bielefeld University, Germany
{cwiljes,cimiano}@cit-ec.uni-bielefeld.de
http://www.sc.cit-ec.uni-bielefeld.de
[2] Central Lab Facilities, CITEC, Bielefeld University, Germany
{flier,tpaulstu}@cit-ec.uni-bielefeld.de
http://www.cit-ec.de/CLF/CentralLabs
[3] Bielefeld University Library, Bielefeld University, Germany
{najko.jahn,johanna.vompras,christian.pietsch}@uni-bielefeld.de
http://www.ub.uni-bielefeld.de/english/

Abstract. For Open Science to be widely adopted, a strong institutional support for scientists will be essential. Bielefeld University and the associated Center of Excellence Cognitive Interaction Technology (CITEC) have developed a platform that enables researchers to manage their publications and the underlying research data in an easy and efficient way. Following a Linked Data approach we integrate this data into a unified linked data store and interlink it with additional data sources from inside the university and outside sources like DBpedia. Based on the existing platform, a concrete case study from the domain of biology is implemented that releases optical motion tracking data of stick insect locomotion. We investigate the cost and usefulness of such a detailed, domain-specific semantic enrichment in order to evaluate whether this approach might be considered for large-scale deployment.

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

1 Motivation

The vision of *Open Science* foresees a scientific publishing environment in which research results are made available and shared openly at all stages of the scientific discovery process. Research results in this sense go beyond the traditional publication of a paper and comprise the release of all important and relevant research artefacts including software, analysis scripts, detailed descriptions of experimental conditions etc. As research becomes more and more data-driven, results can only be independently verified and validated if there is full access to underlying data, processing software, experimental protocols, etc. As such, Open Science promises to increase transparency, integrity and efficiency in science, opening up new avenues for scientific discovery, allowing for data to be

reused in new contexts and fostering the collaboration across disciplines just to name two of many possible benefits. In fact, making the research process more transparent by making all relevant research data publicly available is regarded more and more as a necessity by the research community itself [1] as well as by funding organisations [2].

The Center of Excellence Cognitive Interaction Technology (CITEC[4]), located at Bielefeld University, is a highly interdisciplinary research institute comprising around 250 researchers from disciplines as varied as computer science, biology, physics, linguistics, psychology and sport science. The goal of CITEC is to conduct basic research in cognition while at the same time producing relevant insights and technology that will provide the basis for a better human-machine interaction. The interdisciplinary nature of CITEC calls for a complex communication across institutional and disciplinary borders. Therefore scientists working at CITEC are generally very open to share their research data.

However, the scientists need support and guidance in releasing their data. Data publication is a complicated procedure that involves many different tasks on various levels: organizational, legal and technical. Surveys have repeatedly shown that scientists want to concentrate on their own research questions instead of bothering with technical questions related to data publication [3]. Therefore, in order to be adopted by scientists, the publication of data has to fulfill the following three conditions [4].

1. *easy*: the publication of data should constitute a minimum effort for the scientist
2. *useful*: data publication should not represent a means in itself but offer an immediate and obvious benefit to the scientific community as well as to the scientist him/herself
3. *citable:* data publications have to be citable so they can be referred to within scientific communication and discourse

The goal of our work is to develop an infrastructure that fulfills these three needs at affordable costs of development and operation. The usefulness of a research data management relies on a successful and flexible integration of heterogeneous data from various sources. We will investigate the role Linked Data can play to solve this challenge.

2 Infrastructure

In the following sections we will describe the individual components of the infrastructure at Bielefeld University and CITEC that enable scientists to publish and manage their scientific output. The ultimate goal is to develop a complete ecosystem of services and solutions necessary on the road towards Open Science.

[4] `http://cit-ec.de/`

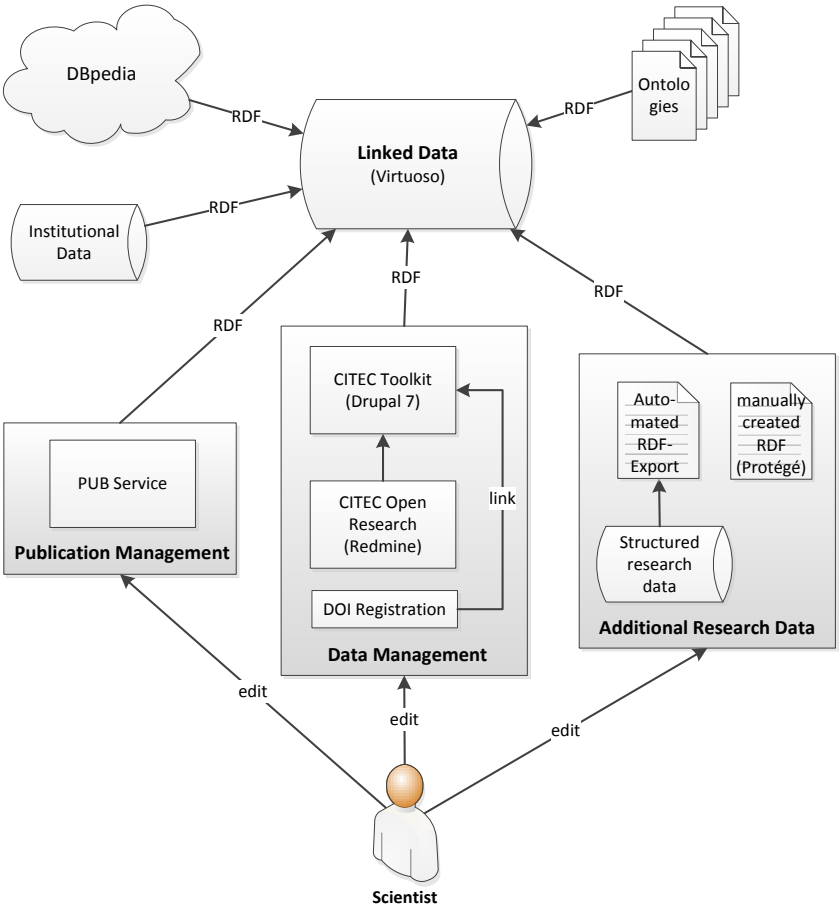

Fig. 1: Open Science infrastructure at Bielefeld University and CITEC with added semantic layer of Linked Data

2.1 Central Services at Bielefeld University

Significant parts of the research infrastructure at Bielefeld University are organised and networked on a university-wide basis.

Publication Management: PUB (Publications at Bielefeld University) provides an overview of publications to which researchers affiliated with Bielefeld University have contributed.[5] This service acts also as an institutional repository allowing to deposit a copy of research publications in accordance with the principle of Open Access to scientific literature. PUB currently has more than 1,000 active users and 35,000 registered publications, of which more than 6,500 provide

[5] `http://pub.uni-bielefeld.de`

a self-archived Open Access fulltext. Highly integrated in the university-wide IT-infrastructure, PUB reuses Bielefeld University's authentication and authorization infrastructure, staff and department profiles as well as grant information to enrich registered publications. Based on this integration, researchers and departments can create dedicated publication profiles to be embedded in the personal or working group homepages[5]. Apart from this, PUB exposes its metadata via OAI-PMH harvesting service[6] and SRU search protocoll[7] in various formats (e.g. Dublin Core, MODS).

PUB has been developed under the umbrella of LibreCat[8] – an open source development project of the university libraries of Lund, Ghent and Bielefeld. Here, common toolsets are shared for user-management (Authentication/Authorization), import/export tools (data cleaning tools, import data sources), metadata management (cataloging tools, indexes, lookup lists) and file management (storage, versioning, fixation).

DOI Registration: For research data to become a research asset comparable to publications, it needs to be persistently available and made citeable like publications. This allows to give credit to the researcher and thereby contribute to the scientist's reputation. In addition it makes sure that the data stays available unchanged over time for later verification and re-use. A solution that has become increasingly popular is assigning digital object identifiers (DOIs) for datasets. A DOI uniquely identifies a dataset, can be resolved via a URL and is accompanied by a set of policies to ensure long-term availability and data integrity.

In 2012, Bielefeld University became a DataCite[9] publication agency and since then started an institutional research data registration service that allows scientists to register their research output. Research groups and faculties interested in a DOI registration for their data have to provide the data itself, the metadata based on the required schema, and the URL of a landing page with, if required, an extended version of the metadata.

2.2 Data Management at CITEC

Research data plays a fundamental role in the scientific process, as it is the basis for developing and testing hypotheses. The internet and the availability of cheap storage space has opened the technical basis for large scale data publication. However, it is not yet common practice for scientists to share their research data for other scientists to verify or use in new contexts.

Researchers frequently produce and make use of various research artifacts, e.g., publications, datasets, experiment specifications, software, etc. Often, these artifacts remain underspecified, lacking important information such as which version of a software or dataset was used for a particular publication. Therefore,

[6] http://www.openarchives.org/pmh/
[7] http://www.loc.gov/standards/sru/
[8] http://librecat.org/
[9] http://datacite.org

reproducing experiments and verifying results sometimes becomes unfeasible. To tackle this issue, CITEC has developed a technical platform and procedures for scientists working at CITEC to publish their research data. This platform consists of two interacting components: the *CITEC Open Research* web platform and the *Cognitive Interaction Toolkit*.

CITEC Open Research Platform: The first component, the *CITEC Open Research* web platform[10], is based on the collaborative development environment *Redmine*[11], a Web application for hosting Open Source projects. The CITEC Open Research platform provides scientists with useful project management features like a wiki, a ticketing system or automated notifications by e-mail. In addition, it offers the possibility to upload and publish digital research data onto a central repository, and provides versioning, integrity checking and long-term support. The integrated wiki offers an easy way to provide documentation.

The central idea behind the CITEC Open Research platform is to offer immediate benefits to scientists so that all relevant research data is automatically accumulated during the whole process of a scientific project and can be released at the end of a project with minimal additional effort. As an additional service, at the finishing stage of a project scientists are supported in finding an adequate licensing model for their data.

Cognitive Interaction Toolkit: The second component, the *Cognitive Interaction Toolkit*[12] [6], is focused on integrating and augmenting existing data from the CITEC Open Research platform and is based on the popular content management system *Drupal*[13]. The core concept of the Cognitive Interaction Toolkit comprises of common research artefact types (e.g., publication, open data set, or software), which can be manually created or imported from external sources like the PUB publication repository.

Inside the Toolkit, researchers can enrich the data by adding relations between datasets, software and connected publications to form aggregates of research artefacts. A linked data representation of aggregates and stand-alone entities is created by the platform semi-automatically based on freely definable vocabularies using Drupal's RDF functionalities. By publishing semantically enriched, functionally relevant aggregates on the web, the Cognitive Interaction Toolkit provides a unified view on research artefacts.

Together, the CITEC Open Research web platform and the Cognitive Interaction Toolkit offer a very flexible, low-cost platform for data management. The Web pages with meta-information about versions of research datasets may act as landing pages resolved via DOIs that allow data to become citable.

[10] `http://openresearch.cit-ec.de/`
[11] `http://redmine.org`
[12] `https://toolkit.cit-ec.uni-bielefeld.de/opendata`
[13] `http://drupal.org`

3 Case Study: Natural Movement Database

We believe that the usefulness of an approach can best be validated by applying it to concrete application cases. Therefore, we implemented a proof of concept scenario to demonstrate the existing infrastructure and to investigate how adding linked data can be beneficial for Open Science. For our case study we chose experiments conducted by the Biological Cybernetics group of CITEC. The motivation and goal of this project has already been presented in detail at SePublica 2012 [7].

Movement is an essential property of animal behaviour. Therefore, understanding movement is an important research question in behavioural neuroscience. The study of movement in biological organisms promises new insights that might be helpful in the creation of artificial systems like robots or embodied agents. At CITEC, movement is being investigated in the context of several research projects. One such project is coordinated by the department of biological cybernetics at Bielefeld University and involves optical motion tracking of stick insects.

The EU project EMICAB[14] conducted at CITEC has set the goal to develop an autonomous hexapod robot [8]. For this, three species of stick insects (*Carausius morosus*, *Aretaon asperrimus*, *Medauroidea extradentata*) are investigated by optical motion tracking. Figure 2 shows a test subject of the species *Aretaon asperrimus* with reflective markers attached. 36 individual test runs of stick insects climbing unrestrained across step obstacles were measured.

Fig. 2: Stick insect with attached markers (used with permission of Volker Dürr)

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

As part of this European project, an open-access Natural Movement Database is being constructed. About 4 hours of recording were created and are to be released as open data towards the end of the project EMICAB. The primary trajectory data measured was transferred into joint-angle-files in MATLAB format that use the test subjects' body model and abstracts from the non-reproducible attachment of the markers on the insects body. The metadata about the experiments was transferred from the files via an import script into a SQL-Database whose schema was custom-designed for this purpose. The overall goal is to store the research data in a structured form that allows publication and re-use in future projects. For this purpose, the process of transforming the primary trajectory data into the relational database has been automated. The raw data will be available as downloads under the DOI `http://dx.doi.org/10.4119/unibi/citec.2013.3`.

4 Adding Linked Data

The technical infrastructure at Bielefeld University and CITEC allows the easy publication of research data alongside publications. However, the data uploaded into the repository is highly heterogeneous both in terms of content and format, and requires intensive documentation to become useful to and interpretable by third parties. As the main goal of an institutional repository of research artifacts is to house a variety of potentially very heterogeneous research objects, several questions need to be addressed:

- How can datasets relevant to a research question be successfully retrieved from a large amount of data?
- How can the data remain interpretable over a long period of time, potentially even after the researcher who created it has left the institution?
- How can the platform be open and flexible enough to allow for the addition of yet unforeseen forms of data in the future?
- How can external sources of data be added to the repository? How can the repository's data be exposed to and used by external services?

A promising approach is to add a semantic layer to the data by representing it as linked data. Linked data can be used to build the connections between research data and the publications that are based on it. Because Linked Data is not bound to a fixed schema, it can be extended to fit project-specific needs. By re-using existing ontologies as widely as possible, connections to external datasources are possible.

Institutional Data: To create an ecosystem of linked data from an institutional repository of research artifacts we need to link to other resources inside our university, i.e. to the scientists who created it, to an organization or project it is associated with, to publications it is related to. Therefore, we set up a knowledge base of linked data that contains data about our university, its institutions and researchers. The URI schema for these resources was defined to

satisfy the requirements of simplicity, stability and manageability as described in [9] and builds on existing identifiers that had already been used inside the IT infrastructure of Bielefeld University. With the VIVO ontology [10], there exists an ontology that can be readily used and covers most of he basic terms needed to describe entities inside a university. The VIVO ontology builds as much as possible on existing, widely-used vocabularies like FOAF, Dublin Core, BIBO and SKOS. Bielefeld University already had a database of its researchers and their organizational affiliations in a relational database that offers an XML interface. Using Extensible Stylesheet Language (XSLT) this data was transformed into an RDF/XML representation. The same approach was applied on the PUB service via its SRU interface, which exposes metadata as MODS-XML.

Data from Cognitive Interaction Toolkit: The Cognitive Interaction Toolkit is based on the CMS Drupal that allows automated generation of linked data. As a first step the Description of a Project (DOAP) ontology[15] was chosen as a vocabulary. This process provides general metadata at minimal cost, but does not go into the details of the specific research. Our goal is to add additional, research-specific information in the form of linked data while still keeping the overall cost manageable.

Data from DBpedia: In addition to these internal resources, connections to external resources are necessary to explicate the content of the data. To account for the heterogeneity of the data and the fact that the content of future datasets cannot be anticipated, a Linked Data repository is needed that is both commonly accepted and covers a spectrum that is broad enough to contain resources from various disciplines. It has been proposed to use *DBpedia*[16], a Linked Data representation of Wikipedia as a crystallization point for the Web of Linked Data [11]. The meaning of URIs is created in a social process like the meaning of words in natural language [12]. Thus the choice of DBpedia/Wikipedia, the largest collaboratively created collection of human knowledge, as a central hub for the emerging web of linked data seems an obvious one. The English version of the DBpedia knowledge base currently describes 3.77 million things and thereby also covers many topics relevant to science. In addition, DBpedia follows Linked Data principles so it has a human readable version for each URI that explains the URIs meaning and it is very well interlinked to other relevant datasets, forming a central hub in the web of Linked Data. Some relevant URIs for our case study are:

```
http://dbpedia.org/resource/Carausius_morosus
http://de.wikipedia.org/wiki/Kleine_Dornschrecke
http://dbpedia.org/resource/Medauroidea_extradentata
http://dbpedia.org/resource/Optical_motion_tracking
```

[15] https://github.com/edumbill/doap/wiki
[16] http://dbpedia.org

It is interesting to notice that one of the three species investigated ("Aretaon asperrimus") has no entry in the english DBpedia, so the German version ("Kleine Dornschrecke") had to be used.

Ontologies: In addition to data from DBpedia, existing, domain-specific ontologies may be used. For the domain of motion tracking, the Ontology for Shape Acquisition and Processing (SAP) [13] is suitable. However, ontology exploration is rather expensive and requires knowledge about the domain as well as an understanding of ontology design. Therefore the cost of exploring or even extending existing ontologies can only be justified in selected cases, i.e. for very valuable data or in cases where a large amount of data from this domain is to be published.

Research Data of Stick Insect Locomotion: As a proof of concept, we exported the motion tracking data from the relational database containing the stick insect locomotion data as RDF/XML using a PERL-Script. The mapping to appropriate RDF-vocabularies is hard coded in the export script. As the data in the database is very subject-specific and fine-grained, only the most important information for interpreting the data has been exported. Listing 1 presents an excerpt of the RDF code generated by exporting the database.

Listing 1: RDF code describing one motion capture experiment (excerpt)

```
<http://info.cit-ec.de/experiment/1> rdf:type dbpedia:Experiment ,
  rdfs:label "Experiment 1" ;
  dc:date "2010-02-17" ;
  dc:title "Step climbing of stick insect Carausius morosus" ;
  sap:hasAcquisitionConditions
    <http://info.cit-ec.de/AcquisitionCondition/1> ;
  sap:hasAcquisitionDevice <http://info.cit-ec.de/equipment/1> ;
  sr:hasSubject dbpedia:Carausius_morosus ;
  citec:hasExperimentalTechnique dbpedia:Optical_motion_tracking ;
  dc:creator <http://info.uni-bielefeld.de/person/18235412> .
```

The resulting RDF was uploaded into a Virtuoso triplestore and exposed by a SPARQL endpoint. A complete set of DBpedia data from the English and the German edition were also imported into the triplestore. The data can be browsed via Virtuososo Faceted Browser Plug-in and graphically explored via Visual Data Web's *Relfinder*[17] [14]. The data, the SPARQL endpoint and the visualization are available online at `http://motion.linked-open-science.org`.

The SPARQL endpoint allows queries that make use of the additional information contained in the internal and external data. Using DBpedia URIs for the test subjects allows us to connect additional data about these species contained in DBpedia and thereby create advanced retrieval methods for research data. For example, Listing 2 displays a SPARQL query that returns all experiments

[17] `http://www.visualdataweb.org/relfinder.php`

about insects, even though the scientist did not explicitly mention that these species are insects.

Listing 2: SPARQL-query: Give me all experiments that investigate insects!

```
PREFIX citec: <http://cit-ec.de/ontology.owl#>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT ?experiment, ?label
WHERE {
  ?experiment dc:subject ?subject .
  ?experiment rdfs:label ?label .
  ?experiment rdf:type dbpedia:Experiment .
  ?subject a dbo:Insect .
}
```

This example illustrates the usefulness of the Linked Data approach: By combining data from two different sources, questions that were not directly foreseen by the providers of the data can be answered. We expect the power of Linked Data to integrate data from various sources will become more apparent as more and more research datasets are released as linked data.

5 Conclusion and Next Steps

The technical infrastructure at Bielefeld University and CITEC allows scientists to publish research data in a central repository and connect it with the corresponding publications that present the results obtained from the data. By assigning DOIs to the datasets research data becomes citable.

We implemented a case study that demonstrates how this infrastructure can act as a Linked Data hub. Linked Data is used to add a semantic layer that enriches the data with additional internal and external data sources, e.g. by linking to DBpedia as a first step, thereby allowing for a more powerful data retrieval. Publication management has already been widely adopted by scientists at Bielefeld University. Adding the research data has recently been installed and is gathering increasing interest and acceptance. Especially the possibility to obtain DOIs for research data, thereby making it citable, increases the attractiveness of data publication to scientists.

Data publication is still a very support-intensive endeavour: DOI registration requires adhering to service level agreements, legal questions need to be addressed, and adding semantic information as Linked Data is still in its infancy, with domain-specific vocabularies still in formation. The challenge for universities is to create an infrastructure and support for scientists that is affordable, easy to use and presents immediate benefits to the scientists.

As case study we presented the natural movement database that collects motion tracking data of stick insects. Publishing the data as static files using the

infrastructure at Bielefeld University has proven to be rather easy and straightforward. However, creating a Linked Data representation of this very domain-specific research data revealed to be rather complex. For future projects, a trade-off between expressiveness and cost needs to be addressed. In general we favour a modular approach that provides for a basic, inexpensive solution but can be enhanced by additional levels of granularity of the semantic enrichment. Which effort is appropriate for the specific data should be decided on a case-by-case basis, depending on the data's nature and value.

We plan to increase the versatility of our system by learning from implementing cases from various disciplines. The goal is to offer a platform that is flexible and powerful enough to adapt to the very heterogeneous requirements from different disciplines, while staying easy to use for the scientists. Our next steps will integrate Linked Data technology more closely with our infrastructure by allowing scientists to directly annotate their data with DBpedia URIs. In addition, we are planning to set up a form-based web front-end that allows scientists to query the linked data stored in the repository's triplestore in an easy and intuitive way.

We believe that successful examples that present a clear benefit to the scientists, both in increasing their scientific reputation and in helping to answer their research questions, will be the best incentive to foster the acceptance of Open Science among scientists.

References

1. Berlin 9 Open Access Conference: Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. http://oa.mpg.de/files/2010/04/berlin_declaration.pdf (2003) [accessed 20-April-2013].
2. Alliance of German Science Organisations: Priority Initiative "Digital Information". http://www.wissenschaftsrat.de/download/archiv/Allianz-digitaleInfo_engl.pdf [accessed 20-April-2013].
3. Feijen, M.: What researchers want - A literature study of researchers' requirements with respect to storage and access to research data. http://www.surffoundation.nl/nl/publicaties/Documents/What_researchers_want.pdf (2011) [accessed 20-April-2013].
4. Smith, V.S.: Data publication: towards a database of everything. BMC research notes **2**(113) (January 2009)
5. Horstmann, W., Jahn, N.: Persönliche Publikationslisten als hochschulweiter Dienst - Eine Bestandsaufnahme. BIBLIOTHEK Forschung und Praxis **34**(2) (2010) 37–45
6. Lier, F., Wrede, S., Siepmann, F., Lütkebohle, I., Paul-Stueve, T., Wachsmuth, S.: Facilitating Research Cooperation through Linking and Sharing of Heterogenous Research Artefacts. In: Proceedings of the 8th International Conference on Semantic Systems. (2012) 157–164
7. Wiljes, C., Cimiano, P.: Linked Data for the Natural Sciences: Two Use Cases in Chemistry and Biology. In: Proceedings of the Workshop on the Semantic Publishing (SePublica 2012). (2012) 48–59

8. Dürr, V., Schmitz, J., Cruse, H.: Behaviour-based modelling of hexapod locomotion: Linking biology and technical application. Arthropod.Struct.Dev. **33**(3) (2004) 237–250
9. Sauermann, L., Cyganiak, R., Völkel, M.: Cool URIs for the semantic web. http://www.dfki.uni-kl.de/dfkidok/publications/TM/07/01/tm-07-01.pdf (February 2007) [accessed 20-April-2013].
10. Conlon, M., Corson-Rikert, J.: VIVO: A Semantic Approach to Scholarly Networking and Discovery (Synthesis Lectures on the Semantic Web). Morgan & Claypool Publishers (2012)
11. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - A crystallization point for the Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web **7**(3) (September 2009) 154–165
12. Halpin, H.: Social Semantics: The Search for Meaning on the Web (Semantic Web and Beyond). Springer (2012)
13. Papaleo, L., Albertoni, R., Pitikakis, M., Robbiano, F., Vasilakis, G., Hassner, T., Moccozet, L., Saleem, W., Tal, A., Veltkamp, R.: Ontology for Shape Acquisition and Processing 4th Version. http://www.aimatshape.net/downloads/public/D1-2-2-1-4th-pdf/download [accessed 20-April-2013].
14. Lohmann, S., Heim, P., Stegemann, T., Ziegler, J.: The relfinder user interface: interactive exploration of relationships between objects of interest. In: Proceedings of the 15th international conference on Intelligent user interfaces. IUI '10, New York, NY, USA, ACM (2010) 421–422

Supplementary Information
The data presented and discussed above, including a SPARQL endpoint that exposes the linked data, is available online at:
http://movement.linked-open-science.org/linked-data.

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

Repurposing Benchmark Corpora for Reconstructing Provenance

Sara Magliacane, Paul Groth

Department of Computer Science
VU University Amsterdam
s.magliacane@vu.nl, p.t.groth@vu.nl

Abstract. Provenance is a critical aspect in evaluating scientific output, yet, it is still often overlooked or not comprehensively produced by practitioners. This incomplete and partial nature of provenance has been recognized in the literature, which has led to the development of new methods for reconstructing missing provenance. Unfortunately, there is currently no agreed upon evaluation framework for testing these methods. Moreover, there is a paucity of datasets that these methods can be applied to. To begin to address this gap, we present a survey of existing benchmark corpora from other computer science communities that could be applied to evaluate provenance reconstruction techniques. The survey identifies, for each corpus, a mapping between the data available and common provenance concepts. In addition to their applicability to provenance reconstruction, we also argue that these corpora could be reused for other tasks pertaining to provenance.

1 Introduction

Data provenance, or the "history of data' [15], is an increasingly important aspect of data management in several settings, forming the foundation of trust, repeatability, attribution and metrics. In scholarly publishing, some aspects of provenance are already tracked manually, traditionally through the mechanisms of authorship and citations. Even though tracking other aspects of provenance would enable a more accurate description of the information flow in science, it is often neglected as a computationally and organizationally intensive task. Moreover, even if the tracking of provenance could be enforced for future publications, how can we include into this vision past and current "legacy" publications?

In most cases, in the absence of actual provenance, we may still prefer to have a plausible reconstruction of what may have happened. One possibility is to automatically generate some hypotheses based on the content of the data and possibly metadata, in a way that resembles forensic investigation, which tries to reconstruct a series of past events based on the available evidence. In this paper, we refer to this task as the problem of provenance reconstruction.

We consider as special cases of provenance reconstruction the following three use cases that are important in scholarly publishing:

– detecting plagiarism, text and multimedia content reuse;

– connecting publications with related data, both research data and other content (blog posts, presentations and videos);
– tracking the evolution of scientific knowledge and discourse through publications and informal communications between scientists;

A number of authors have presented techniques for reconstructing provenance [6,8,7,10,13,14]. However, each approach has been evaluated on different datasets and within different environments. For example, [6] focuses on extracting provenance from newspaper texts whereas [8] uses information from extensive logging within an operating system to create provenance traces. In the context of reconstructing provenance for the scholarly publishing process the related work is evaluated on manually annotated datasets: in particular, [13] describes an approach to reconstruct the provenance of a shared folder containing all the files related to a scientific paper, e.g. TEX files, images and other accessory files, while [14] focuses on the reconstruction of the relationships between a set of papers and clinical guidelines. In some cases, for privacy concerns, the data cannot be made available as is the case with [7]. Because of this heterogeneity, it is difficult to compare the various methods and approaches in a systematic fashion.

Among the openly available provenance datasets, e.g. the ones collected by the ProvBench initiative[1], most focus primarily on provenance generated from computational workflows and not on other environments. An exception is Wikipedia-PROV, a dataset containing the Wikipedia edits provenance graph[2]. Furthermore, often these datasets only provide provenance graphs and not the corresponding information that the provenance refers to. For provenance reconstruction, such information is vital as the reconstruction is based on the content. To evaluate reconstruction methods, one needs to have a gold standard provenance graph and the underlying data from which that gold standard can be built. To begin to address the paucity of datasets, we performed a survey of existing benchmark corpora from other computer science communities that could be applied to evaluate provenance reconstruction techniques. Thus, the paper makes the following contributions:

– a survey of existing benchmark corpora with respect to provenance reconstruction;
– an in-depth analysis of how two of these corpora can be mapped to the W3C PROV model of provenance [11].

More broadly this paper aims to contribute to the ongoing discussion around provenance benchmarks by identifying existing corpora that could be used for benchmark provenance reconstruction approaches. Additionally, it ask the question of how to integrate existing content with next generation publications.

The rest of this paper is organized as follows. We begin by describing our survey methodology and a review the corpora themselves. We then focus on two examples in-depth. Finally, we conclude with some observations about these

[1] https://sites.google.com/site/provbench/
[2] https://github.com/provbench/Wikipedia-PROV

datasets and their applicability to both the specific problem of provenance reconstruction and wider use-cases.

2 Methodology

To collect corpora, we did a focused search concentrating primarily on datasets that have already been used in various computer science evaluation initiatives (e.g TREC). Each corpus was analyzed for its usefulness with respect to provenance and in particular provenance reconstruction. This was done by identifying whether the data could represent information from one or more broad classes of provenance information. The three classes we used are identified below. For each class, we describe how provenance can be concretely expressed using concepts from the W3C PROV data model [16].

- Dependency - a dependency between two objects expressed as the relationship between two prov:Entity objects, e.g. prov:wasRevisionOf or prov:wasDerivedFrom;
- Sequence of operations - a process expressed as a sequence of prov:Activity that connect two prov:Entity objects, expressed through prov:used and prov:wasGeneratedBy relations;
- Authorship - attribution information expressed as the prov:Agent that created the Entity using the prov:wasAttributedTo relation.

These classes reflect the three use-case perspectives on provenance identified by the W3C Provenance Primer [9]: object-oriented, process-oriented and agent-oriented. Thus, these classifications should help guide researchers to useful datasets depending on the perspective their technique is intended for. A key heuristic that we used when deciding whether to incorporate a dataset in the survey was whether it could be used to express not just similarities but dependencies.

3 Survey of existing corpora

There are several available corpora in the Natural Language Processing and Information Retrieval communities. In line with their tasks, most of them provide information about the relevance of entities for a given query or similarity between entities, fewer provide information dependency or influence relationships between entities necessary to act as provenance. Among the existing benchmark corpora that contain provenance-like information most are text-based with only a few containing image and video data.

3.1 Text corpora

Plagiarism detection and text reuse [4] are two related and established fields that can be seen as a special case of reconstructing provenance, especially dependencies between entities. These also play an important role with respect to

scientific literature. Textual entailment can also be seen as a special case of sentence-level provenance. Finally, citation networks provide, what can be seen as, a provenance graph of publications. The following datasets come from these areas.

1. **Corpus Name:** METER corpus [5]
   Availability: Available after registration.[3]
   Background: A journalistic text reuse corpus, consisting of a set of news stories from the major UK news agency and the related news items from nine British newspapers.
   Content: 445 cases of text reuse in 1,716 text documents, annotated by a domain expert in terms of how much the newspaper stories were derived from the agency story and whether there had been some word or phrase reuse. We note that the data was annotated by one, albeit expert, annotator, which could impact upon the accuracy of the information.
   Relationship to Provenance: This corpus can be seen as describing both the dependency and the sequence of operations, reduced to the two basic activities of word reuse and phrase reuse, across the news stories. On the other side, the considered relationships are always from an agency story to a newspaper story, not between agency stories or between newspaper stories.

2. **Corpus Name:** PAN-PC-12 detailed comparison training corpus (an improved version of the PAN-PC-10 [19])
   Availability: Directly available.[4]
   Background: Used in the Plagiarism detection (PAN) 2012 competition[5] in the detailed comparison task.
   Content: The corpus contains 4,210 source documents, derived from the books of Project Gutenberg, and 1,804 "suspicious" documents, where "suspicious" means that they may or may not contained one or more plagiarized passages. In total there are 5,000 plagiarism cases. Each plagiarized passage is annotated with the source passage in the source document. The plagiarism cases were either simulated by crowd-sourcing the rewriting and paraphrasing, or generated artificially through three obfuscation strategies: Random Text Operations (shuffling, removing, inserting or replacing words at random), Semantic Word Variation (replacing word by their synonyms, hyponyms, etc.) and POS-preserving Word Shuffling (shuffling words at random while retaining the original part-of-speech sequence).
   Moreover, there are cases of cross-language plagiarism, which in the past editions of the competition [19] were constructed by applying Google Translate. In PAN-PC-12 they are generated based on the multilingual Europarl corpus [12] by inserting the English version of an originally German or Spanish passage into a Gutenberg book.
   Relationship to Provenance: The released corpus contains information

[3] http://nlp.shef.ac.uk/meter/
[4] http://www.webis.de/research/corpora/corpus-pan-pc-12/pan12/
[5] http://pan.webis.de

on the dependency between entities, in this case paragraphs. This could be improved by tracking the performed operations of the process of automatically generating the corpus. If this was possible, the corpus could also be used as a record of a sequences of operations.

3. **Corpus Name:** Wikipedia co-derivative corpus [2]
   Availability: Available after registration.[6]
   Background: A corpus based on Wikipedia edit history.
   Content: 20,000 documents in four languages (German, English, Hindi and Spanish). For each language, the top 500 most popular Wikipedia articles are retrieved, each with ten revisions.
   Relationship to Provenance: The ten revisions of each article are connected by an edit activity, therefore the corpus contains dependencies between entities. On the other side, the activity is not characterized in more detail, but is just marked as an "edit" operation.

4. **Corpus Name:** PAN-WVC-11 (an improved version on PAN-WVC-10 [18])
   Availability: Directly available.[7]
   Background: Used in the PAN 2011 competition in the Wikipedia vandalism task.
   Content: 29,949 edits on 24,351 Wikipedia articles in three languages (9,985 English edits, 9,990 German edits, and 9,974 Spanish edits), among which 2,813 edits are vandalism edits. The annotated corpus has been crowdsourced using Amazon's Mechanical Turk.
   Relationship to Provenance: This corpus can be thought of as a basic sequence of operations with only one activity, which can be either a legitimate edit or a vandalism edit.

5. **Corpus Name:** PAN-AI-11 training datasets [1]
   Availability: Directly available.[8]
   Background: Used in the Authorship identification task of the PAN 2011 competition, based on a subset of the Enron email dataset.
   Content: More than 12,000 emails written by 118 Enron managers, divided in two subsets:"Large", containing 9337 document by 72 authors, and "Small", containing 3001 documents from 26 authors. The emails were attributed based on the "From: " headers, and multiple emails were reconnected to the same author.
   Relationship to Provenance: The author can be seen as the agent that performs an activity on the document.

[6] http://users.dsic.upv.es/grupos/nle/resources/abc/download-coderiv.html
[7] http://www.uni-weimar.de/cms/medien/webis/research/corpora/corpus-pan-wvc-11.html
[8] http://www.uni-weimar.de/medien/webis/research/events/pan-11/pan11-web/author-identification.html

6. **Corpus Name:** MLaF at FIRE 2011
   Availability: Available after signing the FIRE agreement.[9]
   Background: Used in the Mailing Lists and Forums Track at the FIRE 2011 competition, where the task was the classification of messages from mailing lists and forum discussions in a set of seven types of message.
   Content: 212 132 documents from ubuntu-users (from September 2004 to June 2009) and tugindia (May 2001 - June 2009) mailing lists and on several technical and tech support forums. The corpus maintained the natural causal ordering of the messages in the forums and the threads in the mailing list using the "In-reply-to" fields. Each of the messages is classified as belonging to one or more of seven predetermined categories: e.g. ASK_QUESTION, ASK_CLARIFICATION and SUGGEST_SOLUTION.
   Relationship to Provenance: The dependency relationship is inherent in the structure of the threads, moreover, these categories reflect the activity that generated the messages.

7. **Corpus Name:** RTE-7 [3]
   Availability: Available after signing the Past TAC data agreement[10]. Some older versions (e.g. RTE-3) are available directly.[11]
   Background: Used in the RTE (Recognizing Textual Entailment)[12] challenge at TAC 2011. The main task consisted in determining whether one text fragment is entailed, i.e. can be inferred, from another.
   Content: The corpus contains: a development set of text fragments with 10 topics, 284 hypotheses and 21,420 candidate entailments, of which 1 136 are judged as correct, and a test set with 10 topics, 269 hypotheses and 22,426 candidate entailments, of which 1,308 are judged as correct. The text fragments are based on the TAC 2008 and 2009 Update Summarization Task and the entailment was annotated by three annotators.
   Relationship to Provenance: Textual entailment can be seen as a form of dependency among text fragments. Unfortunately, the activity connecting these fragments cannot be further characterized beyond the entailment.

8. **Corpus Name:** arXiv HEP-PH citation graph from KDD 2003
   Availability: Directly available.[13]
   Background: The articles and citation graph of the high energy physics phenomenology articles uploaded to arXiv between January 1993 and March 2003 from the Citation Prediction task of the 2003 KDD Cup
   Content: 34,546 papers that contain 421,578 references, some of which refer to publications outside of the dataset. The dataset includes the LaTeX source of the main .tex file and several arXiv metadata, like the submission and revision dates, the authors and abstract. Since some of the articles were

[9] https://sites.google.com/site/mlaffire/the-data
[10] http://www.nist.gov/tac/data/past/2011/RTE-7_Main_Task.html
[11] http://pascallin.ecs.soton.ac.uk/Challenges/RTE3/Datasets/
[12] http://www.nist.gov/tac/2011/RTE/
[13] http://www.cs.cornell.edu/projects/kddcup/

older than the submission, it also contains the original publication date.
Relationship to Provenance: The citation networks represent the dependency between publications. The networks not only consider text reuse and paraphrasing, but also textual entailment and summarization. On the other side, if there are any plagiarism cases, they are unlikely to cite the original article so the data may be incomplete with respect to provenance.

3.2 Image corpora

To the best of our knowledge, there is no competition for image reuse or image copy detection, although there is extensive literature on the subject (e.g. see [21] for a comparison of possible approaches). Therefore it was difficult to find publicly available corpora with ground truth annotations that could be repurposed for provenance. The most related competition corpus that we found is used for event detection.

1. **Corpus Name:** Social Event Detection 2012 (SED 2012) dataset [17]
   Availability: Directly available.[14]
   Background: Used at the MediaEval 2012[15] competition in the Social Event Detection task. The task consists in detecting social events and finding clusters of images related to each event. There are three challenges, each related to a specific kind of event, for example the first challenge is "Find technical events that took place in Germany in the test collection".
   Content: 167,332 images captured between the beginning of 2009 and end of 2011 by 4,422 unique Flickr users. For each image there are some metadata available (e.g. time-stamps, tags, 20% of the images have also geotags). The images ewre collected using queries for specific events on the Flickr API.
   Relationship to Provenance: The images of each cluster are all related to the same event, therefore there is a dependency between them.

3.3 Video corpora

Among video retrieval competitions, the most relevant task for our work is the copy detection task. This task has been present at the TRECVid since 2008, but in this paper we describe only the 2011 dataset.

1. **Corpus Name:** CCD at TRECVID 2010[16].
   Availability: Available after signing the TRECVid agreement. Another possibility is to recreate the dataset using the provided tools.
   Background: Used at the TREC Video Retrieval Evaluation (TRECVid)[20] 2010 competition in the content-based copy detection task (CCD). A copy is a segment of video derived from another video using some transformations.

[14] http://mklab.iti.gr/project/sed2012/
[15] http://www.multimediaeval.org/
[16] http://www-nlpir.nist.gov/projects/tv2010/#ccd

Content: The corpus is based on two reference datasets of videos: IACC.1.A, which contains about 8000 Internet Archive videos (MPEG-4 H.264, 50GB, 200 hours) with duration between 10 seconds and 3.5 minutes, and IACC.1.-tv10.training, which contains about 3200 Internet Archive videos (50GB, 200 hours) of around 4 minutes. For most videos the metadata are also available. The queries are constructed from the reference data using specific tools that apply one or more transformations from a known set. This set includes inserting patterns, compression, picture in picture (a video inserted in the front of another video) and post production transformations (e.g. crop, shift, flip).
Relationship to Provenance: There is a dependency between each couple of original and copied video segments that is realized through a sequence of activities, chosen from a known set of transformations.

3.4 Summary of the survey

In Table 1, we present a summary of the surveyed corpora, where each corpus is classified based on the type of data it represents, the operations that are tracked and the information about the authors. In this classification, we did not consider dependencies, because all of the surveyed corpora cover this aspect. An empty cell in the table represents the fact that there is no information regarding to that aspect of provenance. In the case of operations, it means that there are no explicit operations tracked. From Table 1, we can see that for sequences of operations the most promising datasets are PAN-PC-12 [19], MLaF at FIRE 2011 and CCD at TRECVID 2010[20], due to the different categories of operations they capture.

4 Two example corpora

We now look at two corpora in more detail, but the considerations and methods we use could be extended to the other corpora. As a representative of the text corpora, we chose the corpus from the Plagiarism Detection competition (PAN) [19], which provides the most natural and interesting setting for provenance reconstruction. In the view of the increasing multimedia nature of scientific publications, we consider also other forms of content reuse, in particular video content reuse. As a representative of the video corpora, we chose the corpus of TRECVid [20], a well-established competition for video information retrieval and a very good example of sequences of operations reconstruction. For each of these corpora we propose a conversion to the PROV standard, which allows it to be used in a variety of existing applications. Moreover, this conversion enables the connection between these corpora and other provenance datasets.

4.1 Text corpus: PAN-PC-12

We first consider the PAN-PC-12 corpus, which is an updated version of PAN-PC-10 [19]. From this corpus, we consider the detailed comparison corpus, which associates each plagiarized paragraph with the source paragraph. The dataset

Corpus	Type	Operations	Authorship
METER [5]	Newspaper articles (text)	Word reuse, Phrase reuse	-
PAN-PC-12 [19]	Plagiarized books (text)	5 types of plagiarism	-
Wikipedia co-derivative [2]	Wikipedia edits in 4 languages (text)	Edit	-
PAN-WVC-11 [18]	Wikipedia vandalized edits in 3 languages (text)	Edit, vandalization	-
PAN-AI-11 [1]	Emails from 118 authors (text)	-	Email authors
MLaF at FIRE 2011	Mailing lists and forum discussions(text)	7 categories of answer	-
RTE-7 [3]	Text fragments and entailments (text)	-	-
arXiv HEP-PH at KDD 2003	Scientific publications (text)	Cite	Authors
SED-2012	Images about social events (images, tags)	-	-
CCD at TRECVID 2010[20]	Video content reuse examples (video)	10 transformations	-

Table 1: Summary of the survey

contains all the documents in text format and one XML file per document that contains some metadata like author, title and language. Moreover, in the case of suspicious documents, it describes for each plagiarized paragraph the offset and length of the source paragraph, and the source document. The dataset distinguishes several types of plagiarism:

- artificial plagiarism with high/low/no obfuscation;
- translated plagiarism;
- simulated plagiarism with paraphrase (crowd-sourced).

We propose to convert the dataset to a PROV template similar to the one presented in Figure 1. In particular, we model a suspicious document as a collection of several paragraphs, some of which are original and some of which are a result of plagiarism. The plagiarized paragraphs are derived from the original paragraphs through a Plagiarism activity of any of the above-mentioned five types. In addition, each of the original paragraphs is contained in an original document.

4.2 Multimedia corpus: TRECVid 2010

The corpus of the content-based copy detection task of the TREC Video Retrieval Evaluation [20] is a good example multimedia corpus for provenance

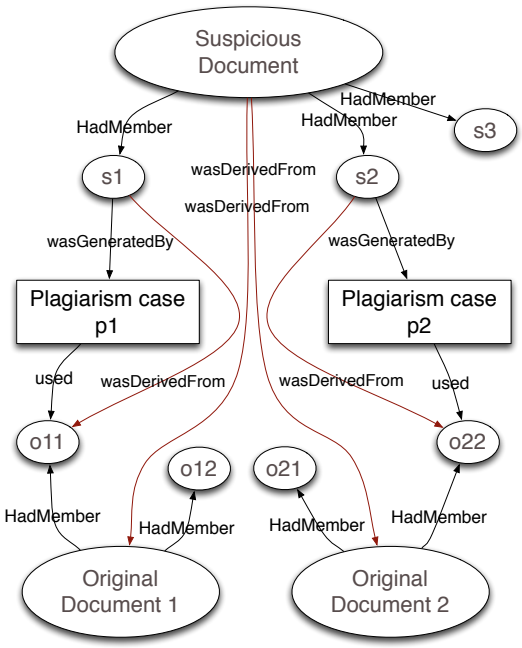

Fig. 1: The PROV template for PAN-PC-12

reconstruction. In the TRECVid terminology, the queries are the plagiarized copies, which are constructed from the two datasets by applying one or more transformation from a known set of ten transformations. Each of the transformations has one or two input videos and several numerical parameters. The transformations can be:

- *basic*, e.g. cam-coding, insertion of a pattern, picture in picture (a video is inserted in the front of another video), blur, crop, shift;
- *composed*, i.e. sequences of three or five basic transformations.

We propose to convert the dataset to a PROV template similar to the one presented in Figure 2. In this case, the generated video is created as an output of the Transformation activity, that can be characterized as one of the types of transformations with certain parameters. The inputs are one or two videos from the reference collection.

5 Analysis & Conclusion

Overall, these corpora provide a good test sets for provenance systems focused on the agent or entity oriented perspectives. However, none of these corpora provide information that can be construed as provenance between different types

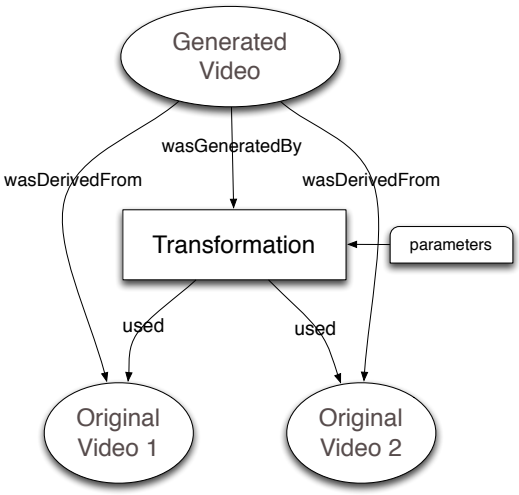

Fig. 2: The PROV template for TRECVid

of entries (e.g. text+image) or as representing long complex and open chains of activities. Additionally, the datasets clearly do not cover all the types of information regularly defined in provenance models, for example, rich semantics about the types of agents or activities within a provenance trace. However, we believe they provide a usable testing environment, in particular, for the reconstruction task. These datasets may also prove useful in systems that want to test the scalability of their provenance infrastructure because of the large amounts of data involved in some of the corpora, for example, TRECVID.

To conclude, in this paper, we provided an overview of existing benchmark corpora that could be used for testing provenance systems and in particular provenance reconstruction methods. We believe that these existing datasets provide a good first step for testing such systems. We hope to provide extracted provenance graphs from a select set of the surveyed datasets. However, going forward, there is a clear need for both manually curated and synthetic provenance-specific benchmarks. The ability to reconstruct provenance will be a key part of integrating existing content into the next generation of scientific publications.

Acknowledgements This publication was supported by the Data2Semantics project in the Dutch national program COMMIT.

References

1. Argamon, S., Juola, P.: Overview of the international authorship identification competition at pan-2011. In: CLEF 2011 (2011)

2. Barrón-Cedeño, A., Eiselt, A., Rosso, P.: Monolingual Text Similarity Measures: A Comparison of Models over Wikipedia Articles Revisions. pp. 29–38
3. Bentivogli, L., Clark, P., Dagan, I.: The seventh pascal recognizing textual entailment challenge. Text Analysis Conference (TAC) 2011 Notebook Proceedings (2011)
4. Broder, A.Z.: On the resemblance and containment of documents. In: In Compression and Complexity of Sequences (SEQUENCES'97 (1997)
5. Clough, P., Gaizauskas, R., Piao, S.: Building and annotating a corpus for the study of journalistic text reuse. LREC 202 (2002)
6. de Nies, T., Coppens, S., van Deursen, D.: Automatic Discovery of High-Level Provenance using Semantic Similarity. IPAW 2012 (2012)
7. Deolalikar, V., Laffitte, H.: Provenance as data mining: combining file system metadata with content analysis. In: First workshop on on Theory and practice of provenance. p. 10. USENIX Association (2009)
8. Frew, J., Metzger, D., Slaughter, P.: Automatic capture and reconstruction of computational provenance. Concurrency and Computation: Practice and Experience 20(5), 485–496 (2008)
9. Gil, Y., Miles, S., Belhajjame, K., Deus, H., Garijo, D., Klyne, G., Missier, P., Soiland-Reyes, S., Zednik, S.: A primer for the prov provenance model (2012), http://www.w3.org/TR/prov-primer/, world Wide Web (W3C)
10. Groth, P., Gil, Y., Magliacane, S.: Automatic Metadata Annotation through Reconstructing Provenance. In: Semantic Web in Provenance Managment workshop (2012)
11. Groth, P., Moreau, L.: http://www.w3.org/TR/prov-overview/
12. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In: Machine Translation Summit X. pp. 79–86. Phuket, Thailand (2005)
13. Magliacane, S.: Reconstructing provenance. The Semantic Web–ISWC 2012 (2012)
14. Magliacane, S., Groth, P.: Towards Reconstructing the Provenance of Clinical Guidelines. In: Proceedings of the 5th International Workshop on Semantic Web Applications and Tools for Life Sciences (SWAT4LS). CEUR Workshop Proceedings, vol. 952. Paris, France (2012)
15. Moreau, L.: The Foundations for Provenance on the Web. Foundations and Trends® in Web Science 2(2-3), 99–241 (2010)
16. Moreau, L., Missier, P.: PROV-DM: The PROV Data Model, http://www.w3.org/TR/prov-dm/
17. Papadopoulos, S., Schinas, E., Mezaris, V., , Troncy, R., Kompatsiaris, I.: Social Event Detection at MediaEval 2012 : Challenges , Dataset and Evaluation. MediaEval 2012 Workshop (2012)
18. Potthast, M.: Crowdsourcing a Wikipedia vandalism corpus. Proceedings of the 33rd international ACM SIGIR 2010 pp. 7–8 (2010)
19. Potthast, M., Stein, B.: An evaluation framework for plagiarism detection. Proceedings of the 23rd International Conference on Computational Linguistics: Posters (2010)
20. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval. pp. 321–330. ACM Press, New York, NY, USA (2006)
21. Thomee, B., Huiskes, M.J., Bakker, E.M., Lew, M.S.: Large scale image copy detection evaluation. In: Lew, M.S., Bimbo, A.D., Bakker, E.M. (eds.) Multimedia Information Retrieval. pp. 59–66. ACM (2008)

Connections across scientific publications based on semantic annotations

Leyla Jael Garcia Castro[1], Rafael Berlanga[1], Dietrich Rebholz-Schuhmann[2], Alexander Garcia[3]

[1] Temporal Knowledge Bases Group, Department of Computer Languages and Systems, Universitat Jaume I, Castello de la Plana, Spain
[2] Institute of Computational Linguistics, University of Zurich, Zurich, Switzerland
[3] Institute for Digital Information & Scientific Communication, College of Communication and Information, Florida State University, Tallahassee, Florida, USA

leylajael@gmail.com, berlanga@lsi.uji.es, rebholz@cl.uzh.ch, alexgarciac@gmail.com

Abstract. The core information from scientific publications is encoded in natural language text and monolithic documents; therefore it is not well integrated with other structured and unstructured data resources. The text format requires additional processing to semantically interlink the publications and to finally reach interoperability of contained data. Data infrastructures such as the Linked Open Data initiative based on the Resource Description Framework support the connectivity of data from scientific publications once the identification of concepts and relations has been achieved, and the content has been interconnected semantically. In this manuscript we produce and analyze the semantic annotations in scientific articles to investigate on the interconnectivity across the articles. In our initial experiment based on articles from PubMed Central we demonstrate the means and the results leading to the interconnectivity using annotations of Medical Subject Headings concepts, Unified Medical Language System terms, and semantic abstractions of relations. We conclude that the different methods would contribute to different types of relatedness between articles that could be later used in recommendation systems based on semantic links across a network of scientific publications.

Keywords: Semantic publication, semantic integration and interoperability, life sciences, semantic annotations, concept recognition.

1 Introduction

Scientific publications have traditionally been the primary means by which scholars communicate their work, e.g., new reporting on hypotheses, methods, results, experiments, etc. [1]. New technologies have introduced changes in the handling of scientific publications; however, the knowledge embedded in such documents remains, to a large extent, poorly exploited and interconnected with other data. The reference section relates scientific articles in an explicit way to other scientific documents, i.e., the prior art [2]. Further relatedness results from shared authors and bibliographic metadata. By contrast, all other connectivity based on the knowledge

representation in the content is underexploited, despite the availability of standardized public resources such as the Medical Subject Headings (MeSH) [3], the Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) [4], and the Unified Medical Language System (UMLS) [5]. These resources would contribute to the construction of knowledge databases facilitating access to semantically normalized information provided by scientific publications.

In this manuscript, we explore on the connections across scientific articles based on their semantic features and annotations. We address the problem of identifying relations between semantic annotations and their relevance for the connectivity between related manuscripts. We examine eleven full-text articles from the open-access subset of PubMed Central and determine which connectivity results from MeSH and UMLS concept annotations. This paper is organized as follows. In Section 2 we introduce our approach while in Section 3 we present the experiment we have carried on, detailing materials, methods, and results. In Section 4 we discuss results and contrast them with related work. Finally in Section 5 we present conclusions and future work.

2 From conceptual features to semantic interconnectivity

We base our approach on the fact that documents do share semantics according to the terminology from the documents. Identifying and annotating terminology has been achieved by different projects, for example the Collaborative Annotation of a Large Biomedical Corpus (CALBC) [6, 7] project. Within CALBC the automatic generation of a large-scale text corpus annotated with biomedical entities, particularly chemical entities, drugs, genes, proteins, diseases, disorders, and species has been studied and the results have been transformed into a triple store [7]. Furthermore, the Nature Publishing Group (NPG) recently released metadata for its publications as Resource Description Framework (RDF) statements; the dataset includes MeSH terms. Finally, the Semantic Enrichment of the Scientific Literature (SESL) [8] project explored the use of semantic web standards and technologies in order to enrich the content of scientific publications: it focused on the integration and interoperability of public and proprietary data resources.

In order to facilitate semantic integration and interoperability for scientific publications, Biotea [9] has built a semantic layer upon the open-access full-text PubMed Central (PMC) articles by transforming the articles into RDF. Biotea also identifies biological entities in the content and abstracts using text-mining and entity-recognition tools, particularly the NCBO Annotator [10] and Whatizit [11, 12]. The identified entities are exposed in RDF as annotations following the model proposed by the Annotation Ontology (AO) [13]. The sets of semantic annotations from the scientific publications facilitate semantic analysis of the unstructured content from the literature.

We augmented the Biotea annotation infrastructure by adding UMLS annotations and by extracting relations involving semantic annotations. In order to identify and semantically categorize these relations, we used several solutions: ReVerb (http://reverb.cs.washington.edu/), a Natural Language Processing (NLP) approach

for relation identification; the Concept Mapping Annotator (CMA) [14]; and a novel semantic-based relation extractor [15]. Both CMA and the relation extractor make use of UMLS, which is one of the most comprehensive knowledge resources in the biomedical domain. Its meta-thesaurus (version 2012AB) covers more than 2.5 million of concepts from over 150 terminological resources, including Medical Subject Headings, NCI Thesaurus, and some others also used for annotations in Biotea. We use the UMLS annotations for the standardization of annotations as well as for the clustering of annotations according to UMLS categories, i.e., the semantic types from the semantic network in UMLS.

In addition and for the future, we propose to include elements of the discourse structure from each manuscript after they have been identified by the SAPIENTA annotator [16]. The relevant Core Scientific Concepts (CoreSC) are labeled as hypothesis, motivation, goal, object, background, method, experiment, model, observation, result and conclusion. Our approach is illustrated in Fig. 1; our main goal is to provide an analytical framework that takes advantage of the semantic features contained in the scientific publications, and focuses on the semantic connections between papers for further information retrieval, recommendation systems and literature-based discovery.

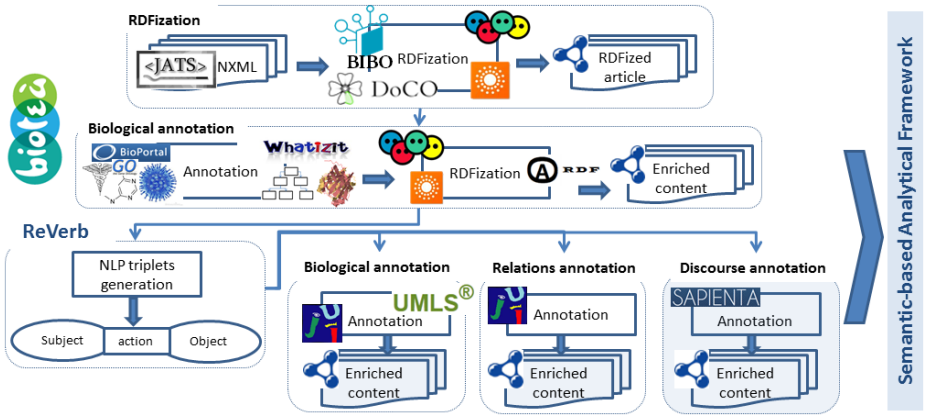

Fig. 1. Our semantic enrichment process. We combine text-mining, entity-recognition, NLP, and semantic techniques in order to provide a semantic layer for scientific publications. The Sapienta components have been shaded, since no results from preliminary experiment will be shown.

3 The augmented Biotea approach

With our analytical framework we have performed an experiment that determines how scientific manuscripts relate to each other based on the co-location of semantic annotations; our analysis relies on concept-based clustering of documents.

3.1 Materials and Methods

From the Biotea SPARQL endpoint (http://biotea.idiginfo.org/query.php), we selected six articles at random from three journals: one from BMC Emergence Medicine, one from Bioinformatics, and four from BMC Biology. All articles satisfy the condition that each one references at least one other manuscript in the endpoint (i.e., ∀x ∃y | x bibo:cites y). In addition to these six articles, we selected five of the referenced articles. Fig. 2 shows the eleven selected articles as well as the SPARQL query which retrieved the initial six documents; Table 1 gives an overview on the selected articles. The process we followed is presented in Fig. 3.

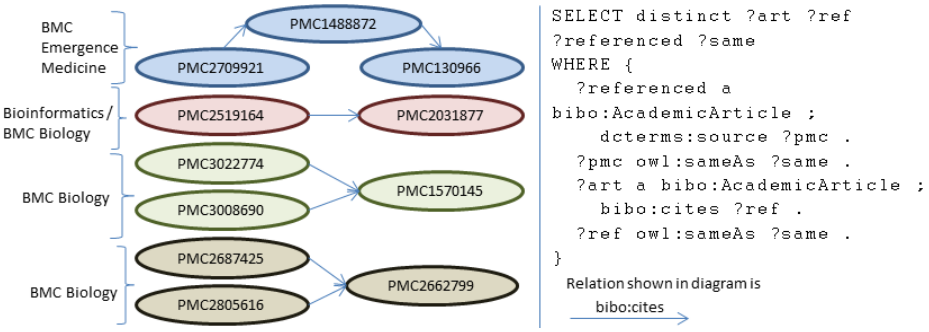

Fig. 2. Selected articles. The SPARQL query included in the figure was used to retrieve the first 100 articles according to the conditions; from them, we selected six referencing at least another one in the subset, and then five of their references. Journals are distinguished by colors.

Table 1. Additional information for selected articles. The five most frequent terms correspond to annotations in Biotea with the highest occurrence in the manuscript.

Articles	Description	Five most frequent terms
PMC 130966	Observational study on patients with suspected acute coronary syndrome (ACS) analyzing characteristics, dispositions, and outcome among patients in order to identify possible improvements in diagnostics.	Patients, ACS, risk, study, symptoms
PMC 1488872	Study on direct hospital costs of chest pain patients in an emergency department (ED)	Patients, cost, ACS, pain, chest
PMC 1570145	Analysis on all genes from sequenced plastid genomes in order to obtain a measure of the overall extent of horizontal gene transfer (HGT) to the plastid	Plastid, genes, HGT, sequence, red
PMC 2031877	Analysis based on the optical projection tomography technique in order to understand how different organ systems and anatomical structures develop throughout the life of	Zebrafish, development, OPT, model, data

	the zebrafish	
PMC 2519164	Review on advances on molecular and cellular microscopic images in bioinformatics, including applications, techniques, tools and resources	Image, analysis, patterns, techniques, data
PMC 2662799	Study baed on full-length sequences of transcripts for *Buchnera aphidicola* and *Acyrthosiphon pisum*, and detailed structural and phylogenetic analyses in order to assess the possibility of lateral gene transfer	Genes, *buchnera*, *ldcA* (gene), *rlpA* (gene), bacteria
PMC 2687425	Commentary on the evolutionary importance of the transfer of genes between host and symbiont	Genes, transfers, genome, host, lateral
PMC 2709921	Evaluation on utility and costs of acute nuclear myocardial perfusion imaging (MPI) in an ED for patients with suspected ACS.	MPI, patients, ACS, cost, study
PMC 2805616	Analysis on integration of non-retroviral ribonucleic acid (RNA) virus genes on fungal hosts, and function of those genes. It uses sequencing across host-virus gene boundaries and phylogenetic analyses of fungal hosts and totivirids	Genes, totivirus, viral, RdRp (RNA polymerase), RNA
PMC 3008690	Summary of two studies related to patterns, processes, and consequences of HGT	Gene, plant, HGT, conversion, gene conversion
PMC 3022774	Analysis on the extent and evolutionary fate of HGT in the parasitic genus *Cuscuta* and a small clade of *Plantago* species aiming to understand details on the mechanics for plant-to-plant HGT	Mitochondrial, transfer, DNA, *plantago*, *atp1* (gene)

In the first step, we retrieved from Biotea the RDF data for the semantic annotations, and selected only those annotations referring to MeSH concepts. We also collected the MeSH terms assigned to the manuscripts in PubMed. In this way we were able to analyze how articles related to each other based on the co-occurrence of MeSH concepts for both datasets, Biotea and PubMed; these first steps correspond to processes 1 and 2 in Fig. 3. From Biotea, we selected the sections and paragraphs, as illustrated by the third step in Fig. 3. We applied ReVerb to the paragraph sections in order to identify sentences that comply with the syntactic form (subject, predicate, object), step 4. As we are interested in the concepts contained in the sentences, we did not discard any sentence at this point of the analysis. For the sentences (see step 5) we applied another annotation tool, called CMA [12]. Similar to the NCBO Annotator and Whatizit, CMA identifies biological entities; furthermore, CMA associates the identified entities with Concept Unique Identifiers (CUIs) from UMLS Meta-thesaurus. Both NCBO Annotator and Whatizit use a dictionary-based text-mining technique while CMA –similar to MetaMap [17]– applies concept classification techniques to stretches of text. For CMA a user may select a threshold to specify the

minimum level of confidence. In our case, we used a low setting to induce high recall. The annotations from CMA contributed in a second analysis towards the relatedness measurements for scientific articles based on the co-occurrence of UMLS terms.

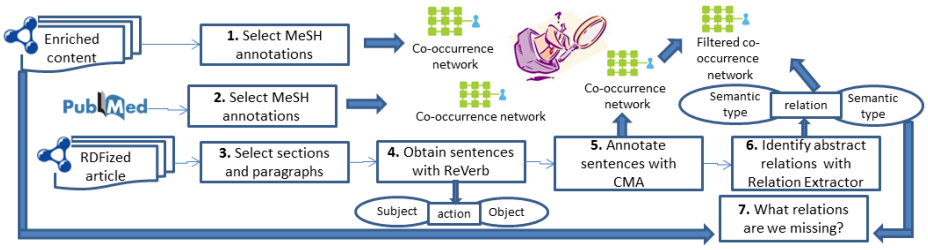

Fig. 3. Our method at a glance.

CMA identifies the subject and object in a sentence, and it is then possible to use [15] in order to identify and categorize the relation between two CUIs (step 6). The semantic relation extractor proposed by [15] extracts the relations between any pair of CUIs in the annotated sentences, resolves synonyms, and produces semantic clusters where the relations are grouped according to synonymous mentions of concepts; in short, it generates abstractions of relations. Such abstract relation form templates where both subject and object make reference to UMLS semantic types; as a result, the template can be applied to any pair of CUIs belonging to the identified semantic types. These identified abstract relations are the basis to the third relatedness analysis.

As we are only receiving relations for those sentences where both subject and object are annotated; we analyzed the annotations provided by Biotea, i.e., the annotations from the NCBO Annotator and Whatizit, for sentences that were not processed by [15]. This way, we could determine the number of relations that went missing (step 7 in Fig. 3). We split the sentences with zero or one recognized CUIs into three subsets: in the first both the subject and the object have been annotated in Biotea, in the second, either subject or object has been annotated, and in the third, all remaining sentences are kept. Below we show the formalization of these subsets in the Formula 1. The first set would tell us exactly how many possible relations we are missing, the second set shows us relations that can be retrieved from ontologies –even when we have identified only one concept in the sentence, other sentences can have ontologically related terms, and the third set contains sentences without enough information for relation extraction.

$$\begin{aligned} S &= \{x \mid x=<\text{subject,action,object}>\} = A \cup B \cup C,\ \emptyset = A \cap B \cap C \\ A &= \{x \mid x=<\text{subject,action,object}> \wedge \text{isAnnotated(subject)} \wedge \text{isAnnotated(object)},\ x \in S\} \\ B &= \{x \mid x=<\text{subject,action,object}> \wedge (\text{isAnnotated(subject)} \veebar \text{isAnnotated(object)}),\ x \in S\} \\ C &= \{x \mid x=<\text{subject,action,object}> \wedge \neg\text{isAnnotated(subject)} \wedge \neg\text{isAnnotated(object)}),\ x \in S\} \end{aligned} \quad (1)$$

3.2 Results

Eleven manuscripts have been annotated with Biotea, CMA, and our semantic relation extractor [15]. In total, the data set comprised 340 paragraphs from 171 sections. We

identified a total of 2088 sentences with ReVerb from which only 1232 had CUIs for both subject and object. From these sentences, a total of 261 abstract relations were extracted. Table 2 gives a summary of our data set.

Table 2. Our working set.

Articles	Sections	Paragraphs	ReVerb Sentences	Analyzed Sentences	Abstract Relations
PMC130966	18	34	150	81	15
PMC1488872	21	30	161	88	27
PMC1570145	26	45	330	172	30
PMC2031877	13	25	164	83	6
PMC2519164	23	51	236	119	3
PMC2662799	21	34	301	177	72
PMC2687425	2	8	73	45	19
PMC2709921	14	31	163	107	25
PMC2805616	10	24	209	122	28
PMC3008690	4	9	56	26	2
PMC3022774	19	49	364	213	34
TOTALS	171	340	2207	1233	261

For the first analysis we examined the connections across the articles based on MeSH concepts. Table 3 presents a summary of the MeSH annotations retrieved from Biotea and PubMed; it also includes the UMLS annotations retrieved with CMA as well as the relations with highest confidence. Annotations from Biotea were retrieved with a SPARQL query while annotations from PubMed were manually gathered. As we were interested in the relatedness between articles, we analyzed the shared annotations that are defined as any concept being referenced as an annotation both in publication *A* and *B*. From the shared annotations we moved to shared concepts, i.e., biological entities associated with a unique entry in a controlled vocabulary.

Table 3. MeSH and UMLS concepts and relation examples in our working set.

Articles	Biotea Mesh	PubMed Mesh	CMA UMLS	Relations with highest confidence
PMC130966	85	Not found	301	Discharged, improved, suitable
PMC1488872	73	Not found	291	Discharged, admitted, defined
PMC1570145	92	8	626	Flanked, matched, extracted
PMC2031877	49	8	302	Examine, prevented, represents
PMC2519164	103	9	466	Begun, attracted, fused
PMC2662799	93	20	799	Encoded, enter, transferred
PMC2687425	28	6	145	Express, functional, reveal
PMC2709921	91	17	330	Discharged, participated, identify
PMC2805616	75	21	361	Encode, integrated, function
PMC3008690	29	8	142	Propose, leads
PMC3022774	79	17	550	Adjacent, converted, enabled

We found 783 shared concepts in Biotea and 33 in PubMed. As the number of shared concepts from Biotea was much higher than the number from PubMed, we selected only concepts with a weight greater than 1.0. The weights varied from 0.04 to 10.41; a total of 67 shared concepts were above the chosen threshold. The weight for shared concepts is defined in Formula 2; as a same concept can be annotated with multiple terms, for instance both "gene" and "genes" could be annotated with the concept MeSH-D005796, we summed up the occurrences by concept rather than term.

$$\text{weight(shared concept C)} = \frac{\left(\frac{\#of\ occurrences\ of\ C}{\#\ of\ sections}\ in\ article\ A + \frac{\#of\ occurrences\ of\ C}{\#\ of\ sections}\ in\ article\ B\right)}{2} \qquad (2)$$

Fig. 4 depicts the connections based on MeSH concepts for Biotea and PubMed. The articles corresponding to BMC Emergence Medicine journal were clustered separately from the rest. This is not clearly visible in the graph that corresponds to annotations from PubMed. We did not find MeSH annotations for the articles PMC130966 and PMC1488772; thus, these two articles together with PMC2709921 are isolated in this graph. In both cases, Biotea and PubMed, PMC2687425 is the most connected article; it has relations to six articles. However, it is not connected to PMC2031877 in Biotea, and to PMC2519164 in PubMed. PMC3022774 is also connected to six articles in Biotea but only to five in PubMed; it is not connected to PMC2031877 in Biotea, and to PMC2031877 and PMC2519164 in PubMed. Surprisingly, although PMC2519164 cites PMC2031877 they are not connected by MeSH concepts.

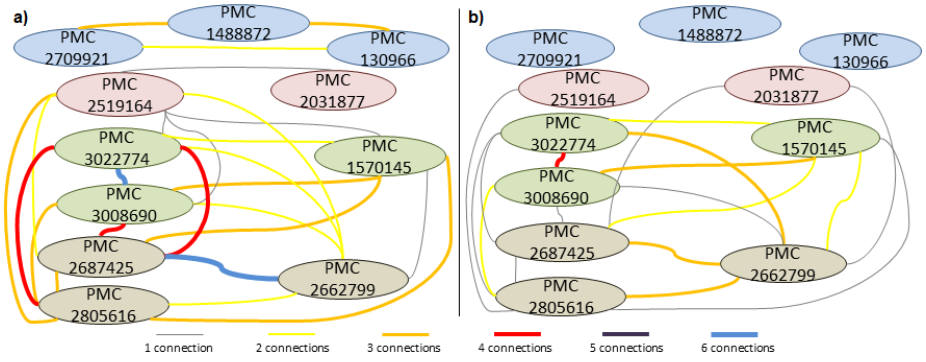

Fig. 4. a) Connections between articles based on MeSH concepts from Biotea. b) Connections between articles based on PubMed MeSH concepts.

Similar to the analysis performed for MeSH terms, we also examined the annotations obtained with CMA. As the threshold was low, we got a large number of terms and concepts; therefore, we also applied the weight formula for CMA annotations. The weights varied from 0.002 to 0.24; we selected the weight 0.1 as cutting point. We found a total of 2343 annotations with CMA covering 2429 different concepts. However, the number could be higher had we annotated the entire

article and not just the sentences identified with ReVerb. Fig. 5 shows the connections according to UMLS concepts from CMA and the extracted relations, i.e., without and with a relation-based filter applied. Similar to the connections from MeSH terms, PMC130966, PMC1488772, and PMC2709921 shaped an independent cluster. However in this occasion PMC1488772 is also connected to PMC2031877, the connections come from the concept "model". The rest of the articles are grouped in a second cluster; there PMC2687425 is connected to all the other articles but PMC2031877, same as it happens in Biotea. Same as it happened in PubMed, PMC3022774 is not connected to either PMC2031877 or PMC2519164. Different as it happened from MeSH connections, this time PMC2687425 is connected to PMC2031877 (indeed the former cites the latter). Fig. 5b shows the same relations but with a relation-based filter applied. From the extracted abstract relations, we chose one "discharge" that takes subjects from the UMLS semantic type T001[LIVB] and objects from T061[PROC]. For PMC1570145 and PMC3008690 no annotation with a semantic type T001 was found.

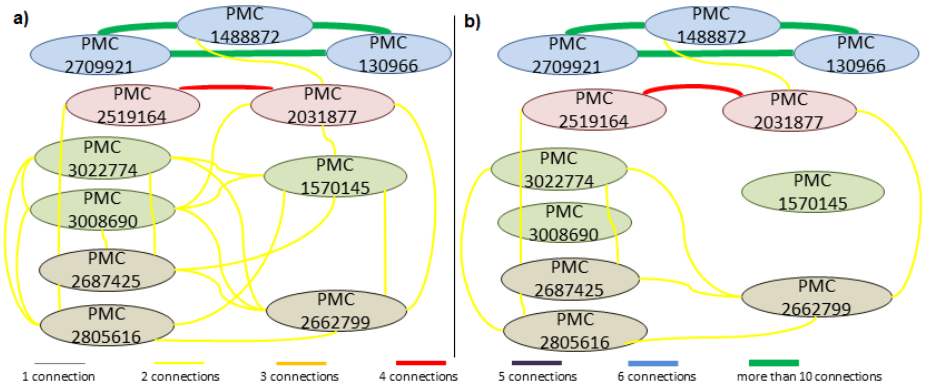

Fig. 5. a) Connections between articles based on UMLS concepts from CMA. b) Connections between articles based on UMLS concepts from CMA with a relation-based filter applied.

Finally, we analyzed the sentences where CUIs were identified only for either the subject or the object, or none of them at all, i.e., for 855 of 2088 retrieved sentences by ReVerb. As illustrated in Formula 1, we partitioned these 975 sentences in three sets: (i) the set A –CUIs for subject and object, with 339 sentences, (ii) the set B – CUIs for either the subject or the object, with 487 sentences; and (iii) the set C –no CUI identified, with 29 sentences. From the relations extractor, we originally got 261 relations from 1233 sentences; corresponding to the 21%. Assuming (i) a linear relation between the number of the sentences with CUIs for both (subject, object) and the abstract relations retrieved, and (ii) only new abstract relations would be retrieved from these 339 new sentences. Then we would be missing about 71 new relations. In relation to the total this is still the 20%. Even when 20% seems to be low, it is important to note that abstract relations actually cover more than one relation as a UMLS semantic type can be applied to multiple CUIs.

4 Discussion

Articles naturally relate to each other via citations; articles sharing citations are considered similar to some extent [2]. Text-based approaches such as term frequency-inverse document frequency and latent semantic analysis have also been used to measure similarity across documents [18]. In a similar vein, cluster-based approaches have also been explored. For instance, Lewis [19] groups related articles by using a keyword-based method followed by a sentence-alignment algorithm that ranks and orders the initial results. Similarly, McSyBi [20] clusters articles according to a set of topics; the information for the creation of topics is gathered from titles and abstracts. Different from Lewis, McSyBi enables the use of MeSH terms or UMLS semantic types in order to modify the clusters so that users can analyze the data from different perspectives.

Unlike these approaches, we are working with a semantically annotated dataset in contrast to plain text articles. Similar to McSyBi, we use MeSH and UMLS concepts in order to calculate relatedness between articles, and explore on the opportunities from semantic annotations of full-text documents. We are reporting connections that later could lead to a semantic-based similarity model for scientific publications. Connections based on MeSH concepts were similar in both cases, Biotea and PubMed. This indicates that it is indeed possible to define a semantic-based approach to measure relatedness across articles. For our working set we only found connections not inferred from PubMed MeSH annotations for those articles without reported MeSH concepts. We cannot conclude yet whether semantic relatedness would be more or less accurate than the relatedness implicit in the related articles suggested in PubMed. However, the similarities in both graphs are a good starting point to extend our relatedness approach to more specific annotations that could introduce a difference; for instance, proteins, genes, diseases, drugs, among others. UMLS connections graph also exhibits similarities with those coming from MeSH terms; therefore, it seems feasible to use other vocabularies, and combining them, in order to find a similarity measure between articles.

Different ways of narrowing the initial connections are possible. For our sample, the relation-based filter applied to the connections did not represent a significant difference. However, it could be improved by using also ontological relations. Even though only connections to two articles were excluded, we still consider that this filtering is a possibility worth exploring further. Rather than using the filtering only for exclusion, it could be also incorporated in the relatedness formula. Although we have explored only the connections across articles, there are other possibilities that can be built on top of semantic dataset for scientific publications as the one provided by Biotea and the extension we propose in this paper. In the biomedical domain, several authors have reported different methods aiming to find hidden relations from semantic annotations. For instance, from MeSH terms it is possible to identify patterns that can be used to find candidates for new associations between drugs and diseases [21]. Similarly, recognizing Gene Ontology terms co-occurring with human gene can be used to discover possible Gene Ontology annotations for those genes [22]. Also, the identification of shared annotations across genes can contribute to identify possible relationships between those genes [23, 24].

5 Conclusions and future work

We have explored how articles connect to each other from a semantic perspective. We have evaluated different concept annotation solutions on full text documents to determine to which extend relatedness can be inferred from such annotations. Such relatedness should facilitate to automatically and semantically integrate literature into an infrastructure of interlinked data elements. Although this semantic-based relatedness project is still in its initial stage, the results from our preliminary experiment are promising. We have found that connections across articles from annotations automatically identified with entity recognition tools, e.g., Whatizit, NCBO Annotator, and CMA, are similar to those connections exhibit based on the PubMed MeSH terms. Having sematic annotations for other vocabularies opens new and interesting possibilities. For instance, it becomes possible to analyze the connections from different perspectives i.e., different vocabularies as well as combinations of them. Additionally, we have also shown the use of relation-based filters in order to narrow the found connections from the co-occurrence of concepts. In our case, we used abstract relations extracted from those sentences where both subject and object were identified by CMA; however, it is also possible to use the relations coming from the ontologies. Different analysis can be performed on the sentences with only one or no biological entities identified; not necessarily about relatedness but also about hidden relations in the plain text.

As part of our future work we have considered to (i) improve the input for ReVerb so we can get more accurate sentences, (ii) use CMA to annotate the entire corpus as it was done in Biotea with the NCBO Annotator and Whatizit, (iii) use relations from the ontologies used to annotated the corpus, (iv) improve our initial weight formula, (v) integrate discourse-based annotations from SAPIENTA, and (vi) formalize a semantic-based method to measure relatedness across scientific publications. The discourse elements provided by SAPIENTA will be used to filter the relations depending on whether or not the participating concepts are related to a particular set of scientific concepts; such set would be define by users.

References

1. Swan, A.: Overview of scholarly communication. In: Jacobs, N. (ed.): Open Access: Key Strategic, Technical and Economic Aspects. Chandos (2006)
2. Hummon, N.P., Dereian, P.: Connectivity in a citation network: The development of DNA theory. Social Networks 11 (1989) 39-63
3. Rogers, F.: Medical subject headings. Bulletin of the Medical Library Association 51 (1963) 114-116
4. Cornet, R., de Keizer, N.: Forty years of SNOMED: a literature review. BMC Medical Informatics and Decision Making 8 Suppl 1 (2008) S2
5. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research 32 (2004) D267-D270
6. Rebholz-Schuhmann, D., Yepes, A., Van Mulligen, E., Kang, N., Kors, J., Milward, D., Corbett, P., Buyko, E., Beisswanger, E., Hahn, U.: CALBC silver standard corpus. Bioinformatics and computational biology 8 (2010) 163-179

7. Croset, S., Grabmüller, C., Li, C., Kavaliauskas, S., Rebholz-Schuhmann, D.: The CALBC RDF Triple Store: retrieval over large literature content. International Workshop on Semantic Web Applications and Tools for the Life Sciences, Berin, Germany (2010)
8. Harrow, I., Filsell, W., Woollard, P., Dix, I., Braxenthaler, M., Gedye, R., Hoole, D., Kidd, R., Wilson, J., Rebholz-Schuhmann, D.: Towards Virtual Knowledge Broker services for semantic integration of life science literature and data sources. Drug Discovery Today in press (2012)
9. Garcia Castro, L.J., McLaughlin, C., Garcia, A.: Biotea: RDFizing PubMed Central in Support for the Paper as an Interface to the Web of Data. Biomedical semantics 4 Suppl 1 (2013) S5
10. Jonquet, C., Shah, N.H., Youn, C.H., Callendar, C., Storey, M.-A., Musen, M.A.: NCBO Annotator: Semantic Annotation of Biomedical Data. International Semantic Web Conference, Poster and Demo session (2009)
11. Rebholz-Schuhmann, D., Arregui, M., Gaudan, M., Kirsch, H., Jimeno, A.: Text processing through Web Services: Calling Whatizit. Bioinformatics 24 (2007) 296-298
12. Kirsch, H., Rebholz-Schuhmann, D.: Distributed modules for text annotation and IE applied to the biomedical domain. International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, Geneva, Switzerland (2004) 50-53
13. 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
14. Berlanga, R., Nebot, V., Jimenez-Ruiz, E.: Semantic annotation of biomedical texts through concept retrieval. Procesamiento de Lenguaje Natural 45 (2010) 247-250
15. Nebot, V., Berlanga, R.: Exploiting semantic annotations for open information extraction: an experience in the biomedical domain. Knowledge and information Systems (2012) 1-25
16. Liakata, M., Saha, S., Dobnik, S., Batchelor, C., Rebholz-Schuhmann, D.: Automatic recognition of conceptualization zones in scientific articles and two life science applications. Bioinformatics 28 (2012) 991-1000
17. Aronson, A.R., Lang, F.-M.: An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association 17 (2010) 229-236
18. Boyack, K.W., Newman, D., Duhon, R.J., Klavans, R., Patek, M., Biberstine, J.R., Schijvenaars, B., Skupin, A., Ma, N., Börner, K.: Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches. PLoS ONE 6 e18029
19. Lewis, J., Ossowski, S., Hicks, J., Errami, M., Garner, H.R.: Text similarity: an alternative way to search MEDLINE. Bioinformatics 22 (2006) 2298-2304
20. Yamamoto, Y., Takagi, T.: Biomedical knowledge navigation by literature clustering. Journal of Biomedical Informatics 40 (2007) 114-130
21. Srinivasan, P., Libbus, B., Sehgal, A.K.: Mining MEDLINE: Postulating a Beneficial Role for Curcumin Longa in Retinal Diseases. In: Hirschman, L., Pustejovsky, J. (eds.): Workshop BioLINK, Linking Biological Literature, Ontologies and Databases at HLT NAACL, Boston, Massachusetts, USA (2004) 33-40
22. Good, B., Su, A.I.: Mining Gene Ontology Annotations From Hyperlinks in the Gene Wiki. Translational Bioinformatics Conference, Washington, D.C. (2011)
23. Saha, B., Hoch, A., Khuller, S., Raschid, L., Zhang, X.-N.: Dense subgraphs with restrictions and applications to gene annotation graphs. 14th Annual international conference on Research in Computational Molecular Biology, Vol. 6044. Springer-Verlag, Lisbon, Portugal (2010) 456-472
24. Thor, A., Anderson, P., Raschid, L., Navlakha, S., Saha, B., Khuller, S., Zhang, X.-N.: Link prediction for annotation graphs using graph summarization. International Conference on the Semantic Web, Vol. 7031. Springer-Verlag, Bonn, Germany (2011) 714-729

Towards the automatic identification of the nature of citations

Angelo Di Iorio[1], Andrea Giovanni Nuzzolese[1,2], and Silvio Peroni[1,2]

[1] Department of Computer Science and Engineering, University of Bologna (Italy)
[2] STLab-ISTC Consiglio Nazionale delle Ricerche (Italy)
diiorio@cs.unibo.it, nuzzoles@cs.unibo.it, essepuntato@cs.unibo.it

Abstract. The reasons why an author cites other publications are varied: an author can cite previous works to gain assistance of some sort in the form of background information, ideas, methods, or to review, critique or refute previous works. The problem is that the best possible way to retrieve the nature of citations is very time consuming: one should read article by article to assign a particular characterisation to each citation. In this paper we propose an algorithm, called *CiTalO*, to infer automatically the function of citations by means of Semantic Web technologies and NLP techniques. We also present some preliminary experiments and discuss some strengths and limitations of this approach.

Keywords: CiTO, CiTalO, OWL, WordNet, citation function, semantic publishing

1 Introduction

The academic community lives on bibliographic citations. First of all, these references are *tools for linking* research. Whenever a researcher writes a paper she/he uses bibliographic references as pointers to related works, to sources of experimental data, to background information, to standards and methods linked to the solution being discussed, and so on. Similarly, citations are *tools for disseminating* research. Not only on academic conferences and journals. Dissemination channels also include publishing platforms on the Web like blogs, wikis, social networks. More recently, semantic publishing platforms are also gaining relevance [15]: they support users in expressing semantic and machine-readable information. From a different perspective, citations are *tools for exploring* research. The network of citations is a source of rich information for scholars and can be used to create new and interesting ways of browsing data. A great amount of research is also being carried on sophisticated visualisers of networks of citations and powerful interfaces allowing users to filter, search and aggregate data. Finally, citations are *tools for evaluating* research. Quantitative metrics on bibliographic references, for instance, are commonly used for measuring the importance of a journal (e.g. the *impact factor*) or the scientific productivity of an author (e.g. the *h-index*).

This work begins with the basic assumption that all these activities can be radically improved by exploiting the actual nature of citations. Let us consider citations as means for evaluating research. Could a paper that is cited many times with negative reviews be given a high score? Could a paper containing several citations of the same research group be given the same score of a paper with heterogeneous citations? How can a paper cited as plagiarism be ranked? These questions can be answered by looking at the nature of the citations, not only their existence. On top of such characterisation, it will also be possible to automatically analyse the pertinence of documents to some research areas, to discover research trends and the structure of communities, to build sophisticated recommenders and qualitative research indicators, and so on.

There are in fact ontologies for describing the nature of citations in scientific research articles and other scholarly works. In the Semantic Web community, the most prominent one is *CiTO* (*Citation Typing Ontology*)[3] [12]. CiTO is written in OWL and is connected to other works in the area of semantic publishing. It is then a very good basis for implementing sophisticated services and for integrating citational data with linked data silos.

The goal of this paper is to present a novel approach to automatically annotate citations with properties defined in CiTO. We present an algorithm and its implementation, called *CiTalO* (from merging the words ***CiTO*** and ***al*** *gorithm*), that takes as input a sentence containing a reference to a bibliographic entity and infers the function of that citation by exploiting Semantic Web technologies and Natural Language Processing (NLP) techniques. The tool is available online at `http://wit.istc.cnr.it:8080/tools/citalo`.

We also present some preliminary tests on a small collection of documents, that confirmed some strengths and weaknesses of such approach. The research direction looks very promising and the CiTalO infrastructure is flexible and extensible. We plan to extend the current set of heuristics and matching rules for a wide practical application of the method.

The paper is then structured as follows. In Section 2 we introduce previous works on classification of citations. In Section 3 we describe our algorithm introducing its structure and presenting the technologies (NLP tools, sentiment analysis procedures, OWL ontologies) we used to develop it. In Section 4 we present the outcome of the algorithm run upon some scientific documents and we discuss those results in Section 5. Finally, in Section 6, we conclude the paper sketching out some future works.

2 Related works

The automatic analysis of networks of citations is gaining importance in the research community. Copestake *et al.* [4] present an infrastructure called SciBorg that allows one to automatically extract semantic characterisations of scientific texts. In particular, they developed a module for discourse and citation analysis based on the approach proposed by Teufel *et al.* [17] called *Argumentative*

[3] CiTO: http://purl.org/spar/cito.

Zoning (*AZ*). AZ provides a procedural mechanism to annotate sentences of an article according to one out of seven classes of a given annotation scheme (i.e. *background*, *own*, *aim*, *textual*, *contrast*, *basis* and *other*), thus interpreting the intended authors' motivation behind scientific content and citations.

Teufel *et al.* [18] [19] study the *function* of citations – that they define as "author's reason for citing a given paper" – and provide a categorisation of possible citation functions organised in twelve classes, in turn clustered in *Negative*, *Neutral* and *Positive* rhetorical functions. In addition, they describe the outcomes of some tests involving hundreds of article in computational linguistics (stored as XML files), several human annotators and a machine learning approach for the automatic annotation of citation functions. Their approach is quite promising; however the agreement between human annotators (i.e. K = 0.72) is still higher than the one between the human annotators and the machine learning approach (i.e. K = 0.57).

Jorg [9] introduces an analysis of the ACL Anthology Networks[4] and identifies one hundred fifty *cue verbs*, i.e. verbs usually used to carry important information about the nature of citations: *based on*, *outperform*, *focus on*, *extend*, etc. She maps cue verbs to classes of citation functions according to the classification provided by Moravcsik *et al.* [10] and makes the bases to the development of a formal citation ontology. This works actually represent one of the sources of inspiration of *CiTO* (the *Citation Typing Ontology*) developed by Peroni *et al.* [12], which is an ontology that permits the motivations of an author when referring to another document to be captured and described by using Semantic Web technologies such as RDF and OWL.

Closely related to the annotation of citation functions, Athar [1] proposes a sentiment-analysis approach to citations, so as to identify whether a particular act of citing was done with positive (e.g. praising a previous work on a certain topic) or negative intentions (e.g. criticising the results obtained through a particular method). Starting from empirical results Athar *et al.* [2] expand the above study and show how the correct sentiment (in particular, a negative sentiment) of a particular citation usually does not emerge from the citation sentence – i.e. the sentence that contains the actual pointer to the bibliographic reference of the cited paper. Rather, it actually becomes evident in the last part of the *context window*[5] [14] in consideration.

Hou *et al.* [8] use an alternative approach to understand the importance (seen as a form of positive connotation/sentiment) of citations: the citation counting in text. Paraphrasing the authors, the idea is that the more a paper is cited within a text, the more its scientific contribution is significative.

[4] ACL Anthology Network: http://clair.eecs.umich.edu/aan/index.php.

[5] The *context window* [14] of a citation is a chain of sentences implicitly referring to the citation itself, which usually starts from the citation sentence and involves few more subsequent sentences where that citation is still implicit [3].

3 Our approach

In this section, we introduce CiTalO, a tool that infers the function of citations by combining techniques of ontology learning from natural language, sentiment-analysis, word-sense disambiguation, and ontology mapping. These techniques are applied in a pipeline whose input is the textual context containing the citation and the output is a one or more properties of CiTO [12].

The overall CiTalO schema is shown in Fig. 1. It was inspired by Gangemi *et al.*'s work [7], in which a similar pipeline was used with good results for automatically typing DBpedia resources by analysing corresponding Wikipedia abstracts. Five steps (described below) compose the architecture, and each one is implemented as a pluggable OSGi component [11] over a Pipeline Manager that coordinates the process.

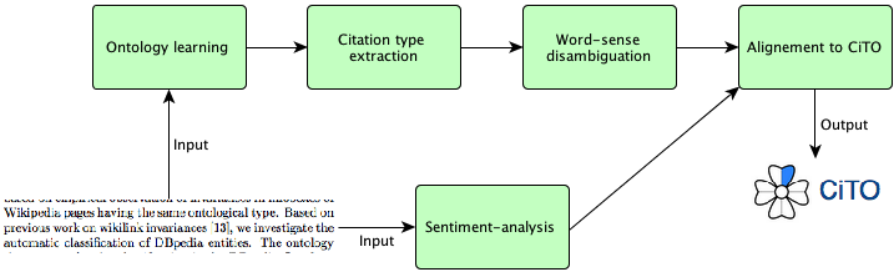

Fig. 1. Pipeline used by CiTalO. The input is the textual context in which the citation appears and the output is a set of properties of the CiTO ontology.

In order to detail the components of CiTalO we will discuss how the algorithm works on the following sample sentence: "It extends the research outlined in earlier work X.", where "X" is the cited work.

Sentiment-analysis to gather the polarity of the citational function. The aim of the sentiment-analysis in our context is to capture the sentiment polarity emerging from the text in which the citation is included. The importance of this step derives from the classification of CiTO properties according to three different polarities, i.e., positive, neuter and negative. This means that being able to recognise the polarity behind the citation would restrict the set of possible target properties of CiTO to match. We are currently using AlchemyAPI[6], a suite of sentiment-analysis and NLP tools that exposes its services through HTTP REST interfaces. The output returned by this component with respect to our example is a positive polarity.

Ontology extraction from the textual context of the citation. The first mandatory step of CiTalO consists of deriving a logical representation of

[6] AlchemyAPI: http://www.alchemyapi.com.

the sentence containing the citation. The ontology extraction is performed by using FRED [13], a tool for ontology learning based on discourse representation theory, frames and ontology design patterns. Such an approach follows the one proposed by Gangemi *et al.* [7], which exploited FRED for automatically typing DBpedia entities. The transformation of the sentence into a logical form allows us to recognise graph-based heuristics in order to detect possible types of functions of the citation. The output of FRED on our example is shown in Fig. 2. FRED recognises two events, i.e., *Outline* and *Extend*,and the cited work *X* is typed as *EarlierWork* that is subclass of *Work*.

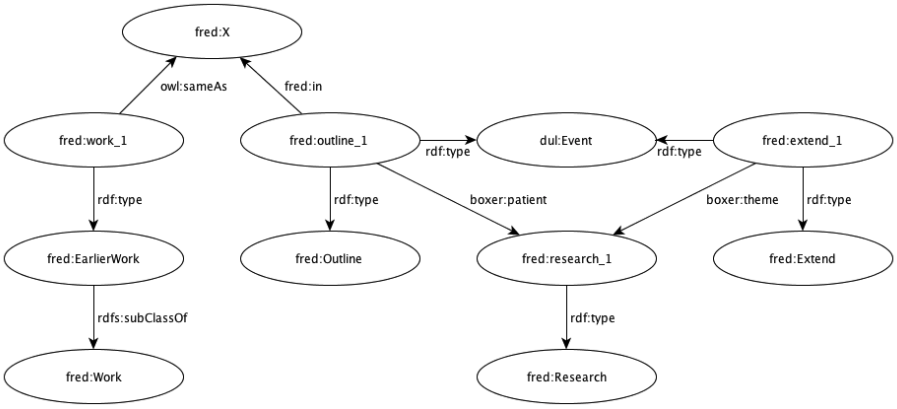

Fig. 2. FRED result for "It extends the research outlined in earlier work X".

Citation type extraction through pattern matching. The second step consists of extracting candidate types for the citation, by looking for patterns in the FRED result. In order to collect these types we have designed ten graph-based heuristics and we have implemented them as SPARQL queries. The pattern matcher tries to apply all the patterns, which are namely:

```
SELECT ?type WHERE {?subj ?prop fred:X ; a ?type}
SELECT ?type WHERE {?subj ?prop fred:X ; a ?typeTmp .
  ?typeTmp rdfs:subClassOf+ ?type}
SELECT ?type WHERE {?subj a dul:Event , ?type .
  FILTER(?type != dul:Event)}
SELECT ?type WHERE {?subj a dul:Event , ?typeTmp .
  ?typeTmp rdfs:subClassOf+ ?type.FILTER(?type != dul:Event)}
SELECT ?type WHERE {?subj a dul:Event ;
  boxer:theme ?theme . ?theme a ?type}
SELECT ?type WHERE {?subj a dul:Event ; boxer:theme ?theme .
  ?theme a ?typeTmp . ?typeTmp rdfs:subClassOf+ ?type}
SELECT ?type WHERE {?subj a dul:Event ;
  boxer:patient ?patient . ?patient a ?type}
SELECT ?type WHERE {?subj a dul:Event ; boxer:patient ?pat .
```

```
  ?pat a ?typeTmp . ?typeTmp rdfs:subClassOf+ ?type}
SELECT ?type WHERE {?subj a dul:Event ; boxer:patient ?pat .
  ?pat ?prop ?any . ?any a ?type}
SELECT ?type WHERE {?subj a dul:Event ; boxer:patient ?pat .
  ?pat ?prop ?any . ?any a ?typeTmp .
  ?typeTmp rdfs:subClassOf+ ?type}
```

Applying these patterns to the example the following candidate types are found: *Outline, Extend, EarlierWork, Work, and Research.* The current set of patterns is quite simple and incomplete. We are investigating new patterns and we are continuously updating the catalogue.

Word-sense disambiguation. In order to gather the sense of candidate types we need a word-sense disambiguator. For this purpose we used IMS [20], a tool based on linear support vector machines. The disambiguation is performed with respect to OntoWordNet [6] (the OWL version of WordNet) and produces a list of synsets for each candidate type. The following disambiguations are returned on our example: (i) *Extend* is disambiguated as `own:synset-prolong-verb-1`, (ii) *Outline* as `own:synset-delineate-verb-3`, (iii) *Research* as `own:synset-research-noun-1`, (iv) *EarlierWork* and *Work* as `own:synset-work-noun-1`.

The output of this step can also be extended by adding *proximal synsets,* i.e. synsets that are not directly returned by IMS but whose meaning is close to those found while disambiguating. To do so, we use the RDF graph of proximality introduced in [7].

Alignment to CiTO. The final step consists of assigning CiTO types to citations. We use two ontologies for this purpose: *CiTOFunctions* and *CiTO2Wordnet*. The CiTOFunctions ontology[7] classifies each CiTO property according to its factual and positive/neutral/negative rhetorical functions, using the classification proposed by Peroni *et al.* [12].

CiTO2Wordnet[8] maps all the CiTO properties defining citations with the appropriate Wordnet synsets (as expressed in OntoWordNet). This ontology was built in three steps:

- *identification step.* We identified all the Wordnet synsets related to each of the thirty-eight sub-properties of *cites* according to the verbs and nouns used in property labels (i.e. *rdfs:label*) and comments (i.e. *rdfs:comment*) – for instance, the synsets *credit#1, accredit#3, credit#3, credit#4* refers to the property *credits*;
- *filtering step.* For each CiTO property, we filtered out all those synsets of which the *gloss*[9] is not aligned with the natural language description of the property in consideration – for instance, the synset *credit#3* was filtered out since the gloss "accounting: enter as credit" means something radically different to the CiTO property description "the citing entity acknowledges contributions made by the cited entity";

[7] CiTOFunctions: http://www.essepuntato.it/2013/03/cito-functions.
[8] CiTO2Wordnet ontology: http://www.essepuntato.it/2013/03/cito2wordnet.
[9] In Wordnet, the *gloss* of a synset is its natural language description.

– *formalisation step.* finally, we linked each CiTO property to the related synsets through the property *skos:closeMatch.* An example in Turtle is: `cito:credits skos:closeMatch synset:credit-verb-1`.

The final alignment to CiTO is performed through a SPARQL CONSTRUCT query that uses the output of the previous steps, the polarity gathered from the sentiment-analysis phase, OntoWordNet and the two ontologies just described. In the case of empty alignments, the CiTO property *citesForInformation* is returned as base case. In the example, the property *extends* is assigned to the citation.

4 Testing and evaluation

The test consisted of comparing the results of CiTalO with a human classification of the citations. The test bed we used for our experiments includes some scientific papers (written in English) encoded in XML DocBook, containing citations of different types. The papers were chosen among those published in the proceedings of the Balisage Conference Series. In particular, we automatically extracted citation sentences, through an XSLT document[10], from all the papers published in the seventh volume of Balisage Proceedings, which are freely available online[11]. For our test, we took into account only those papers for which the XSLT transform retrieved at least one citation (i.e. 18 papers written by different authors). The total number of citations retrieved was 377, for a mean of 20.94 citations per paper. Notice that the XSLT transform was quite simple at that stage. It basically extracted the *citation sentence* around a citation (i.e. the sentence in which that citation is explicitly used), preparing data for the actual CiTalO pipeline.

We first filtered all the citation sentences from the selected articles, and then we annotated them manually using the CiTO properties. Since the annotation of citation functions is actually an hard problem to address – it requires an interpretation of author intentions – we mark only the citations that are accompanied by verbs (extends, discusses, etc.) and/or other grammatical structures (uses method in, uses data from, etc.) carrying explicitly a particular citation function. We considered that rule as a strict guideline as also suggested by Teufel *et al.* [18].

We marked 106 citations of out the 377 originally retrieved, obtaining at least one representative citation for each of the 18 paper used (with a mean of 5.89 citations per paper). We used 21 CiTO properties out of 38 to annotate all these citations, as shown in Table 1.

Interesting similarities can be found between such a classification and the results of Teufel *et al.* [19]. In this paper, the neutral category *Neut* was used for the majority of annotations by humans; similarly the most neutral CiTO property, *citesForInformation*, was the most prevalent function in our dataset too. The second most used property was *usedMethodIn* in both analyses.

[10] Available at http://www.essepuntato.it/2013/sepublica/xslt.
[11] Proceedings of Balisage 2011: http://balisage.net/Proceedings/vol7/cover.html.

Table 1. The way we marked the citations within the 18 Balisage papers.

# Citations	CITO property
53	citesForInformation
15	usesMethodIn
12	usesConclusionsFrom
11	obtainsBackgroundFrom
8	discusses
4	citesAsRelated, extends, includesQuotationFrom, citesAsDataSource, obtainsSupportFrom
< 4	credits, critiques, useConclusionsFrom, citesAsAuthority, usesDataFrom, supports, updates, includesExcerptFrom, includeQuotationForm, citesAsRecommendedReading, corrects

We run CiTalO on these data (i.e. 106 citations in total) and compared results with our previous analysis[12]. We also tested eight different configurations of CiTalO, corresponding to all possible combinations of three options:

- activating or deactivating the sentiment-analysis module;
- applying or not the proximal synsets[13] to the word-disambiguation output;
- using the CiTO2Wordnet ontology as described in Section 3, or an extended version that also includes all the discarded synsets during the filtering step.

The number of *true positives* (*TP*), *false positives* (*FP*) and *false negatives* (FN) obtained comparing CiTalO outcomes with our annotations are shown in Table 2.

We calculated the precision – i.e. `TP / (TP + FP)` – and the recall – i.e. `TP / (TP + FN)` – obtained by using each configuration. As shown in Fig. 3, *Filtered* and *Filtered+Sentiment* had the best precision (i.e. 0.348) and the second recall (i.e. 0.443), while *All* and *All+Sentiment* had the second precision (i.e. 0.313) and the best recall (i.e. 0.491).

There is no configuration that emerges as the absolutely best one from these data. They rather suggest an hybrid approach that also takes into account some of the discarded synsets. It is evident that the worst configurations were those that took into account all the proximal synsets. It looks that the more synsets CiTalO uses, the less the citation functions retrieved conform to humans' annotations.

[12] All the source materials we used for the test is available online at http://www.essepuntato.it/2013/sepublica/test. Note that a comparative evaluation with other approaches, such as Teufel's, was not feasible at this stage since input data and output categories were heterogeneous and were not directly comparable.

[13] We used the same the RDF graph of proximal synsets introduced in [7].

Table 2. The number of true positives, false positives and false negatives returned by running CiTalO with the eight different configurations.

Configuration	TP	FP	FN
Filtered (with or without Sentiment)	47	88	59
Filtered + Proximity	40	137	66
Filtered + Proximity + Sentiment	41	136	65
All (with or without Sentiment)	52	114	54
All + Proximity (with or without Sentiment)	45	174	64

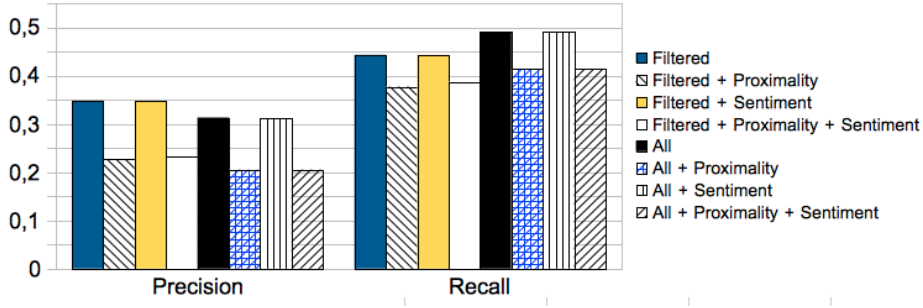

Fig. 3. Precision and recall according to the different configurations used.

In general, the values of precision and recall of our experiments are quite low. However, our preliminary tests aimed at defining a baseline for future developments of our approach, more than a definitive evaluation of CiTalO effectiveness.

5 Limitations and future research directions

In this section we discuss some limitations and possible improvements of CiTalO outlined by the tests and that we plan to address in future releases of the tool.

Coverage of CiTO properties. The manual annotation process highlighted that CiTO properties do not cover all the citation scenarios addressed in the experiment. For instance, let us consider the following sentence from [16]: "*We speculate that some Goddag-based structure analogous to the multi-coloured trees of [Jagadish et al. 2004] may be a useful solution*".

The verb *speculate* used above is very specific and refers to synsets that are not included in the mapping defined in the CiTO2Wordnet ontology. This kind of citation is not explicitly mentioned in CiTO neither. The same happens for citations – usually they are introduced by modal verbs – that suggest a work as *potential solution* for an issue related to the paper in consideration, for instance (again from [16]): "*Mechanisms like Trojan Horse markup ([DeRose 2004], [Bauman 2005]) can be used to serialize discontinuous elements*".

What is needed here is an accurate analysis of citations in papers so as to suggest some extensions to CiTO itself. Towards this direction, a good starting point is to use Jorg's previous work on cue verbs [9], where she listed one hundred-fifty verbs that are typically used in citations within scientific articles.

Noise of proximity synsets. The diagram in Fig. 3 clearly shows that using proximity synsets decreased both precision and recall. One would expect, on the other hand, that a larger set of synsets produced better results.

This depends on the number of *citesAsInformation* retrieved by CiTalO (remind that *citesAsInformation* is assigned when no further CiTO property is identified). Let us consider the case of *Filtered*: *citesForInformation* was assigned correctly 42 times out of 47 occurrences[14], while using *Filtered+Proximity* the same property was detected only 31 times and other more specific CiTO properties were assigned instead. The problem is that those assignments are not correct, as they derive from proximal synsets that are actually too far from the ones being processed in CiTO2Wordnet. These synsets should not be considered or, at least, should be given less importance than others that are closer to the ones in CiTO2Wordnet. For future releases of CiTalO, in fact, we plan to use proximal synsets distance in order to reduce such a noise.

Matching synsets and compound-word properties. The current CiTalO alignment between synsets and CiTO properties does not work properly with properties described by compound words, such as *useMethodIn*. In fact, CiTalO returns a match whether one of the synsets of the compound words matches with a CiTO property. For instance, let us consider the following sentence (from [16]): "*Later versions of the TEI Guidelines [ACH/ACL/ALLC 1994] define more powerful methods of encoding discontinuity*".

CiTalO returns the property *usesMethodIn* since one of the related synsets of that property, i.e. *synset:method-noun-1*, was actually found. This output is not correct, since that property should be returned only if there exists evidence that the current work *uses* (a term that is actually missed from that sentence) a particular *method* from another article, while here it seems not to be the case. Future version of CiTalO must take into account these scenarios too.

Identification of the context window of citations. In our experiments, we always used the citation sentence as input of CiTalO. However, as previously noticed by Athar *et al.* [2], the actual intended sentiment and motivation of a citation is not always present in the citation sentence. It may be explicit in some other sentences close to the citation sentence and can refer implicitly to the cited work (through authors' names, project's name, pronouns, etc.). The identification of the right citational *context window* [14] is a complex issue that should be addressed to improve the effectiveness of CiTalO.

Identification of implicit citations. The identification of *implicit citations* [3] is another issue related to the one being discussed. Let us consider some sentences of a paragraph from [16]: "*XCONCUR and similar mechanisms*

[14] The other citation functions retrieved are: *citesAsRecommendedReading*, *usesDataFrom*, *citesAsDataSource*, *extends* and *usesMethodIn* – all of them used just one time within the true positive set.

[Hilbert/Schonefeld/Witt 2005] already incorporate the containment/dominance distinction to a certain degree. [...] And like non-concurrent XML, XCONCUR has no conception of discontinuous elements".

While in the first sentence, it seems that the authors want to praise with a positive connotation the work done by others (i.e. XCONCUR), in the latter sentence they criticise them. The "XCONCUR" in the latter sentence actually represents an implicit citation of the reference contained in the former sentence and, in this case, delimits also the context window of the citation itself. Detecting such scenarios is a further refinement that can improve CiTalO results.

Using rhetoric structures. According to Teufel *et al.* [18], recognising implicit citations and context windows "is often not informative enough for the searcher to infer the relation" of citations. Further information can be given by also identifying the rhetorical function of the entire paragraph or section in which the citation appears. For instance, all the references in the "related works" section are usually used to indicate related articles (i.e. *citesAsRelated*) to the topic under consideration, while citations in the introduction present background information (i.e. *obtainsBackgroundFrom*) of the field in which the work described in the article is placed. We are thinking to apply existing techniques of automatic recognition of document structures, e.g. that proposed by Di Iorio *et al.* [5], to retrieve the rhetoric function of sections in scientific articles and integrate such analysis with CiTalO.

6 Conclusions

The implementation of CiTalO is still at an early stage; current experiments are admittedly not enough to fully validate this approach. However, the overall approach is very open to incremental refinements. The goal of this work, in fact, was to build such a modular architecture, to perform some exploratory experiments and to identify issues and possible developments of our approach. We are currently working to include a mechanism for the automatic identification of *context windows* of citations given an input article and to improve *patterns' matching* phases in CiTalO. In addition, we plan to perform exhaustive tests with a larger set of documents and users.

References

1. Athar, A. (2011). Sentiment Analysis of Citations using Sentence Structure-Based Features. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: 81-87.
2. Athar, A., Teufel, S. (2012). Context-Enhanced Citation Sentiment Detection. In Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics 2012: 597-601.
3. Athar, A., Teufel, S. (2012). Detection of implicit citations for sentiment detection. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: 18-26.

4. Copestake, A., Corbett, P., Murray-Rust, P., Rupp, C. J., Siddharthan, A., Teufel, S., Waldron, B. (2006). An architecture for language processing for scientific text. In Proceedings of the UK e-Science All Hands Meeting 2006.
5. Di Iorio, A., Peroni, S., Poggi, F., Vitali, F. (2012). A first approach to the automatic recognition of structural patterns in XML documents. In Proceedings of the 2012 ACM symposium on Document Engineering: 85-94. DOI: 10.1145/2361354.2361374
6. Gangemi, A., Navigli, R., Velardi, P. (2003). The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet. In Proceedings of CoopIS/DOA/ODBASE 2003: 820–838. DOI: 10.1007/978-3-540-39964-3_52
7. Gangemi, A., Nuzzolese, A.G., Presutti, V., Draicchio, F., Musetti, A., Ciancarini, P. (2012). Automatic Typing of DBpedia Entities. In Proceedings of the 11th International Semantic Web Conference: 65-81. DOI: 10.1007/978-3-642-35176-1_5
8. Hou, W., Li, M., Niu, D. (2011). Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution. In BioEssays, 33 (10): 724-727. DOI: 10.1002/bies.201100067
9. Jorg, B. (2008). Towards the Nature of Citations. In Poster Proceedings of the 5th International Conference on Formal Ontology in Information Systems.
10. Moravcsik, M. J., Murugesan, P. (1975). Some Results on the Function and Quality of Citations. In Social Studies of Science, 5 (1): 86-92.
11. OSGi Alliance (2003). OSGi service platform, release 3. IOS Press, Inc.
12. Peroni, S., Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. In Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 17 (December 2012): 33-43. DOI: 10.1016/j.websem.2012.08.001
13. Presutti, V., Draicchio, F., Gangemi, A. (2012). Knowledge extraction based on discourse representation theory and linguistic frames. In Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management: 114-129. DOI: 10.1007/978-3-642-33876-2_12
14. Qazvinian, V., Radev, D. R. (2010). Identifying non-explicit citing sentences for citation-based summarization. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: 555-564.
15. Shotton, D. (2009). Semantic publishing: the coming revolution in scientific journal publishing. In Learned Publishing, 22 (2): 85-94. DOI: 10. 1087/2009202
16. Sperberg-McQueen, C. M., Huitfeldt, C. (2008). Markup Discontinued: Discontinuity in TexMecs, Goddag structures, and rabbit/duck grammars. In Proceedings of Balisage: The Markup Conference 2008. DOI: 10.4242/BalisageVol1.Sperberg-McQueen01
17. Teufel, S., Carletta, J., Moens, M. (1999). An annotation scheme for discourse-level argumentation in research articles. In Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics: 110-117.
18. Teufel, S., Siddharthan, A., Tidhar, D. (2006). Automatic classification of citation function. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing: 103-110.
19. Teufel, S., Siddharthan, A., Tidhar, D. (2009). An annotation scheme for citation function. In Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue: 80-87.
20. Zhong, Z., Ng, H. T. (2010). It Makes Sense: A wide-coverage word sense disambiguation system for free text. In Proceedings of the ACL 2010 System Demonstrations: 78-83.

How Reliable is Your workflow: Monitoring Decay in Scholarly Publications

José Manuel Gómez-Pérez[1], Esteban García-Cuesta[1], Jun Zhao[2], Aleix Garrido[1], José Enrique Ruiz[3], Graham Klyne[2]

[1] Intelligent Software Components (iSOCO), Spain.
[2] University of Oxford, UK.
[3] Instituto de Astrofísica de Andalucía, Spain.
jmgomez@isoco.com

Abstract. Scientific workflows play an important role in computational research, as the essential artifacts for communicating the methods used to produce the research findings. We are witnessing a growing number of efforts of treating workflows as first-class artifacts for sharing and exchanging actual scientific knowledge, either as part of scholarly articles or as stand-alone objects. However, workflows are not born to be reliable, which can seriously damage their reusability and trustworthiness as knowledge exchange instruments. Scientific workflows are commonly subject to decaying, which consequently undermines their reliability. In this paper, we propose the hypothesis that reliability of workflows can be notably improved by advocating scientists to preserve a minimal set of information that is essential to assist the interpretations of these workflows and hence improve their reproducibility and reusability. By measuring and monitoring the completeness and stability of this information over time, we are then able to indicate the reliability of scientific workflows, which is critical for establishing trustworthy reuse of these important scientific artifacts and supporting the claims in related publications.

1 Introduction

Scientific workflows are well-known means to encode scientific knowledge and experimental know-how. By providing explicit and actionable representations of scientific methods, workflows capture such knowledge and support scientific development in a number of critical ways, including the validation of experimental results and the development of new experiments based on the reuse and repurposing of existing workflows. Therefore, scientific workflows are valuable scholarly objects themselves and play an important role for sharing, exchanging, and reusing scientific methods. In fact we are witnessing a growing trend of treating workflows as first-class artifacts, for exchanging and transferring actual scholarly findings, either as part of scholarly articles or as stand-alone objects, as illustrated by popular public workflow repositories like myExperiment [6] and CrowdLabs [15].

Reliability of workflows, i.e. the claimed capability of a workflow, is key to its reuse and as the instrument for knowledge exchange. However, reliability of a

workflow can hardly be guaranteed throughout its life time. Scientific workflows are commonly subject to a decayed or reduced ability to be executed or repeated, largely due to the volatility of the external resources that are required for their executions. This is what we call *workflow decay* [20]. Workflow definitions, which record the processes/services used or the data processed, clearly cannot capture all the information required to preserve the original capability of the workflows. For example, information about the originator of a workflow is one key piece of information to establish trust for a workflow. A workflow created by a reputable research group or researcher is expected to be more reliable. But this attribution and credit information about workflows may be difficult to address without additional information like provenance metadata about the author.

In order to support these needs for enhancing the reliability of workflows, we propose the adoption of workflow-centric research objects [2] to encapsulate additional information along with workflows, as one single information unit. Such information, structured in the form of semantic annotations following standards like the Annotation Ontology [5], OAI-ORE[1] and PROV-O [2], describes the operations performed by the workflow, provides details on authors, versions or citations, and links to other resources, such as the provenance of the results obtained by executing the workflow, input and output datasets or execution examples. Research objects consequently provide a comprehensive view of the experiment, support the publication of experimental results, enable inspection, and contain the information required for the evaluation of the health of a workflow.

In this paper we propose the hypothesis that workflow reliability can be notably improved by preserving a minimal set of essential information along with workflows themselves. This requires a systematic understanding of the causes to workflow decay and hence the set of information to be preserved to prevent or reduce decay. In [20] we produced a classification of causes to workflow decay by systematically analysing a corpus of Taverna workflows selected from the popular public workflow repository myExperiment.org. Based on our analysis, we identified the minimal set of information to be associated in a workflow to reduce its decay and proposed a minimal information model (Minim) to represent these information as quality requirements that must be satisfied by a research object.

This paper takes a step forward in this direction. Research objects enable scientists to safeguard their workflows against decay by defining and evaluating against a minimal set of requirements that must be satisfied. However, there is a lack of indicators that provide third party scientists with the necessary information to decide whether an existing workflow is reliable or not. Workflows are commonly subject to changes over their life span. On one hand this is due to the nature of knowledge evolution. Workflows are often working scholarly objects that are part of a larger scientific investigation. As scientific understandings develop, workflow designs must be updated accordingly. On the other hand,

[1] http://www.openarchives.org/ore/1.0/toc.html
[2] http://www.w3.org/TR/prov-o

given the volatile external context that a workflow is built upon, throughout the investigation a workflow may be subject to various changes, to deal with for example, updates of external data formats, data access methods, etc. Our method must consider both these internal and external changes when helping the scientists to judge the reliability of a workflow: a workflow that works at the time of inspection cannot be quickly concluded as reliable; while one which does not cannot be simply dismissed as unreliable.

In [20] we introduced the notion of completeness of a research object, i.e., the degree by which a research object contains all the required resources necessary for a purpose (e.g., workflow runnability). In this paper we introduce a new metric, stability, which measures the ability of a workflow to preserve its overall completeness state throughout a given time period. We combine the stability measure with the completeness measure in order to compute the reliability of a workflow. Stability extends the scope of the analysis from a particular point in time to a given time period. Parameters like the impact of the information added or removed from the research object and of the decay suffered throughout its history are taken into account for the computation. In this paper we also present an analytic tool that enables scientists and other stakeholders to visualize these metrics and have a better understanding of the evolution of workflow reliability over time.

The remainder of the paper is structured as follows. Section 2 provides an account of related work relevant for the evaluation of workflow reliability. In section 3 we motivate the need for using completeness and stability measures to establish workflow reliability. We then present an outline of our approach in section 4 and describe our implementation in section 5. In section 6, we illustrate the application of our approach using a case study. Finally, section 7 concludes by summarizing our main contributions and outlining future work.

2 Related Work

Our discussion spans through different areas relevant for scholarly communication dealing with: the modelling of aggregation structure as the basis of new ways of publication and the definition of metrics that assess the information being communicated is conserved free of decay throughout time.

While [14] argued in favor of the use of a small amount of semantics as a necessary step forward in scholarly publication, research objects were conceived to extend traditional publication mechanisms [1] and take us beyond the pdf [4] by aggregating essential resources related to experiment results along with publications. This includes not only the data used but also methods applied to produce and analyze those data. The notion of using aggregation to promote reproducibility and accessibility of research has been studied elsewhere, including the Open Archives Initiative Object Reuse and Exchange Specification (OAI-ORE) [18], the Scientific Publication Packages (SPP)[13], and the Scientific Knowledge Objects [8]. Nano-publication [12] is another approach of

supporting accessible research by publishing key results of an investigation as concise statements.

Along those lines, an important part of the role of workflow-centric research objects as publication objects is to ensure that the scientific method encoded by a workflow is actually reproducible, therefore providing evidence that the results claimed by the authors actually hold. This has a strong impact in the reuse of workflow-based experiments [9] and is closely related to the goal of myExperiment packs [17], which aggregate elements such as workflows, documents and datasets together, following Web 2.0 and Linked Data principles, in order to support communication and reuse of scientific methods.

In order to enhance the trustworthiness of these ROs we associate them with a list of explicitly defined requirements that they must satisfy and we use this list to evaluate their completeness, i.e. the quality of the ROs with respect to a set of given criteria. This is built upon the idea of a Minimum Information Model (MIM) [7], which provides an encoding of these requirements in OWL[3] and supports reasoning with them. Also related to this is work on information quality in the Web of Data [3] and, more specific to the e-science domain, [16], which focuses on preventing experimental work from being contaminated with poor quality data resulting from inaccurate experiments.

Finally, approaches like [10] aim at validating the execution of specific workflows by checking the provenance of their execution against high level abstractions which act as semantic overlays and allow validating the correct behaviour of the workflow. Complementary work from the field of monitoring and analysis of web-scale service based applications like [11] aims at understanding and analyzing service-oriented applications and eventually detecting and preventing potential misbehaviour.

3 Motivation

To illustrate the need of assessing the reliability of a workflow as a fundamental indicator for reuse, we use an example research object based on a workflow from myExperiment[4] in the Astronomy domain, used to calculate distances, magnitutes and luminosities of galaxies.

In this scenario, Bob has a list of several tens of galaxies that have been observed by members of his group during the last years. He is trying to find a workflow which performs queries on services from the International Virtual Observatory[5] (VO) in order to gather additional complementary physical properties for his galaxies. Related to the tag *extragalactic*, Bob finds a promising workflow in a research object published by Alice. He reads its description and finds some similarities to his problem. He also has a list of galaxies and would like to query several web services to access their physical properties, though not the same as those in Alice's case, and perform similar calculations on them. Bob

[3] http://www.w3.org/2004/OWL/
[4] http://www.myexperiment.org/workflows/2560
[5] http://www.ivoa.net

inspects some of the components of Alice's research object in order to better understand it and to find out what parts he could reuse. Several of the input datasets provided in the research object are interesting, as well as their related information and semantic descriptions.

After successfully running the workflow, Bob finally feels confident that Alice's workflow is a perfect candidate for reuse in his own work. However, a deeper analysis of its recent history could prove otherwise:

1. The workflow evolution history shows that one of the web services changed the format of the input data when adopting ObsTAP VO[6] standards for multidata querying. As a consequence the workflow broke, and authors had to replace the format of the input dataset.
2. This dataset was also used in a script for calculating derived properties. The modification of the format of the dataset had consequences in the script, which also had to be updated. Bob thinks this may be very easily prone to errors.
3. Later on, another web service became unavailable during a certain time, which turned out that the service provider (in fact Bob's research institution) forgot to renew the domain and the service was down during two days. The same happened to the input data, since they were hosted in the same institution. Bob would prefer now to use his own input dataset, and not to rely on these ones.
4. This was not the only time the workflow experienced decay due to problems with its web services. Recent replacement of networkign infrastructure (optic fiber and routing hardware) had caused connectivity glitches in the same institution, which is the provider of the web service and input datasets. Bob wonders if he could find another web service to replace this one. He needs his workflow working regularly, since it continuously looks for upgraded data for his statistical study.
5. Finally, very recently a data provider modified the output format of the responses from HTML to VOTable[7] format, in order to be VO compliant and achieve data interoperability. This caused one of the scripts to fail and required the authors to fix it in order to deal with VOTable format instead of proprietary HTML format. Bob thinks this is another potential cause for having scripts behaving differently and not providing good results.

In summary, even though the workflow currently seems to work well, Bob does not feel totally confident about its stability. The analysis shows that trustworthy reuse depends not only on the degree to which the properties of a particular workflow and its corresponding research object are preserved but also on their history. This is especially true for scientists who, like Bob, think a particular workflow can be interesting for them but lack the information about its recent performance. Workflows which can be executed at a particular point in

[6] http://www.ivoa.net/Documents/ObsCore
[7] http://www.ivoa.net/Documents/VOTable

time might decay and become unrunnable in the future if they depend on brittle service or data infrastructure, especially when these belong to third party institutions. Likewise, if they are subject to frequent changes by their author and contributors, the probability that some error is introduced also increases. Therefore, we introduce the concept of workflow stability as a means to consider its recent history an background to evaluate its reliability.

4 Approach

We understand *reliability* as a measure of the confidence that a scientist can have in a particular workflow to preserve its capability to execute correctly and produce the expected results. A reliable workflow is expected not only to be free of decay at the moment of being inspected but also in general throughout its life span. Consequently, in order to establish the reliability of a workflow it becomes necessary to assess to what extent it is complete with respect to a number of requirements and how stable it has been with respect to such requirements historically. Therefore, we propose *completeness* (already introduced in [20]) and *stability* as the key dimensions to evaluate workflow reliability. Figure 1 schematically depicts the reliability concept as a three-tiered compound on top of completeness and stability along time.

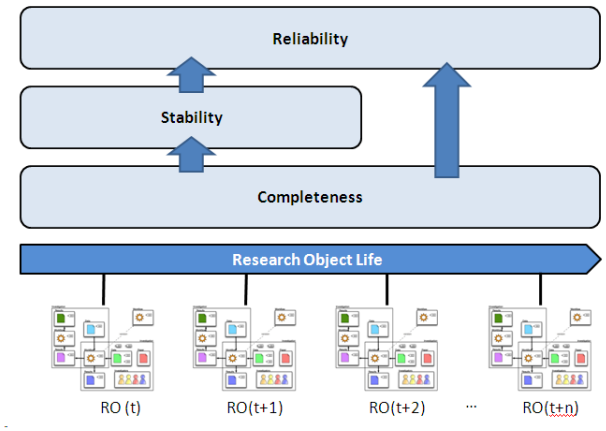

Fig. 1. Layered Components of Reliability Measurement

Following the figure, the next sections define each dimension and the relations between them, from completeness to stability and finally reliability.

4.1 Completeness

The completeness dimension evaluates the extent to which a workflow satisfies a number of requirements specified in the form of a checklist following the Minim OWL ontology[8]. Such requirements can be of two main types: compulsory (must) or recommendable (should). In order to be runnable and reproducible all the *must* requirements associated to a workflow need to be satisfied while *should* requirements propose a more relaxed kind of constraint. An example of the former is that all the web services invoked by the workflow be available and accessible (two of the main causes of workflow decay), while the presence of user annotations describing the experiment would illustrate the former.

Since *must* requirements have a strong impact we have defined two thresholds: a) a lower bound β_l which establishes the maximum value that the completeness score can have in case it does not satisfy all *must* requirements, and b) an upper bound β_u which establishes the maximum value that the completeness score can have given that it satisfies all *should* and *must* requirements. Both β_l and β_u are parameterizable and can be configured on a case by case basis.

Therefore if at least a *must* requirement fails the completeness score is in the lower band $[0-\beta_l]$ and otherwise in the upper band $[0-\beta_u]$. Once identified the band, we define a normalized value of the completeness score as:

$$completeness\_score(RO,t) = f(RO_{(t)}, requirements, type) = \alpha\frac{nSReq(RO_{(t)},must)}{nReq(must)} + (1-\alpha)\frac{nSReq(RO_{(t)},should)}{nReq(should)} \in [0,1],$$

where t is the point in time considered, RO the research object that contains the workflow being evaluated, $requirements$ the specific set of requirements defined within the RO for a specific purpose, $type \in \{must, should\}$ the category of the requirement, $\alpha \in [0,1]$ is a control value to weight the different types of requirements, $nSReq$ the number of satisfied requirements, and $nReq$ the total number of requirements for the specified type. This definition of the completeness score guarantees the following properties:

- The maximum value possible if a *must* requirement fails is defined by the lower bound β_l.
- The maximum value possible if all requirements are satisfied is defined by the upper bound $\beta_u = 1$.

4.2 Stability

The stability of a workflow measures the ability of a workflow to preserve its properties through time. The evaluation of this dimension provides the needed information to scientists like Bob the astronomer in order to know how stable the workflow has been in the past in terms of completeness fluctuation and therefore to gain some insight as to how predictable its behavior can be in the near future. We define the stability score as follows:

[8] http://purl.org/net/minim/minim#

$$stability\_score(RO, t) = 1 - std(completeness\_score(RO, \Delta t)) \in [0.5, 1],$$

where completeness_score is the measurement of completeness in time t and Δt is the period of time before t used for evaluation of the standard deviation. The stability score has the following properties:

- It reaches its minimum value when there are severe changes over the resources of a workflow for the period of time Δt, meaning that the completeness score is continuously switching from its minimum value of zero (bad completeness) to its maximum of one (good completeness). This minimum value is therefore associated to unstable workflows.
- It has its maximum value when there are not any changes over a period of time Δt, meaning that the completeness score does not change over that time period. This maximum value is therefore associated to stable workflows.
- Its convergence means that the future behavior of the workflow can be predictable and therefore potentially reusable by interested scientists.

4.3 Reliability

The reliability of a workflow measures its ability for converging towards a scenario free of decay, i.e. complete and stable through time. Therefore, we combine both measures completeness and stability in order to provide some insight into the behavior of the workflow and its expected reliability in the future. We define the reliability score as:
$reliability\_score(RO, t) = completeness\_score(RO, t) * stability\_score(RO, t) \in [0, 1]$,

where RO is the research object, and t the current time under study. The reliability score has the following properties:

- It has a minimum value of 0 when the completeness score is also minimum.
- It has a maximum value of 1 when the completeness score is maximum and the RO has been stable during the period of time Δt
- A high value of the measure is desirable, meaning that the completeness is high and also that it is stable and hence predictable.

5 Implementation: RO-Monitoring Tool

In this section we describe our developed RO-Monitoring tool, which implements the criteria of completeness, stability, and reliability as formulated in section 4.

Our monitoring tool provides functionalities for time-based computation of the completeness, stability and reliability scores of an RO via a Restful API[9], and stores the results as additional metadata within the RO, as shown in the following sample excerpt of RO metadata in RDF turtle notation. The complete sample

[9] http://sandbox.wf4ever-project.org/decayMonitoring/rest/getAnalytics

RO including this excerpt and the rest of encapsulated metadata, following the RO ontologies[10], and materials can be found in the RO digital library[11] of the Wf4Ever project. The monitoring trace of the RO can be visualized through the RO-Monitoring tool[12].

```
@prefix rdfs:    <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd:     <http://www.w3.org/2001/XMLSchema#> .
@prefix owl:     <http://www.w3.org/2002/07/owl#> .
@prefix rdf:     <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

<http://sandbox.wf4ever-project.org/rodl/ROs/Pack387/>
      <http://purl.org/wf4ever/rovalues#completeness>   1.0 ;
      <http://purl.org/wf4ever/rovalues#reliability>    0.869 ;
      <http://purl.org/wf4ever/rovalues#stability>      0.869 .
```

The resulting information allows producing analytics of the evolution of these metrics over time, as shown in Figure 2. The tool is closely based on the Restful checklist service, previously presented in [20], which evaluates the completeness of a workflow-oriented research object according to quality requirements expressed using the Minim OWL ontology. In addition to this monitoring service, the RO monitoring tool also provides a web-based user interface, using JavaScript and jQuery libraries. Through this interface, users can inspect the values of these metrics for an RO in time, compare differences between any two time points and, more importantly, gain access to an explanation of these changes. Therefore it is possible for users to have a quick overview of who has changed what in an RO, and the impact of such actions in terms of reliability.

The RO-Monitoring service makes use of the Research Object Evolution Ontology (roevo[13]) to provide explanations to any changes occurred in a time span, e.g. a sudden drop in the reliability score. Built upon the latest PROV-O standards, the roevo ontology enables the representation of the different stages of the RO life-cycle, their dependencies, changes and versions. Using the RO evolution traces together with the reliability scores, we can offer end users meaningful explanations for helping them to interpret the reliability variations, like the number of changes, its type, or the author of those changes.

6 Monitoring RO Decay in Practice

This section shows how our RO-Monitoring tool works in practice with the astronomy case study described in section 3. Figure 2 shows the results produced by the RO-Monitoring tool which visualizes the reliability trace of an astronomy workflow based on the completeness scores computed by daily evaluations, and the stability and reliability scores computed on top of them.

[10] http://www.wf4ever-project.org/research-object-model
[11] http://sandbox.wf4ever-project.org/rodl/ROs/Pack387
[12] http://sandbox.wf4ever-project.org/decayMonitoring/monitorReliability.html?id=lt
[13] http://purl.org/wf4ever/roevo

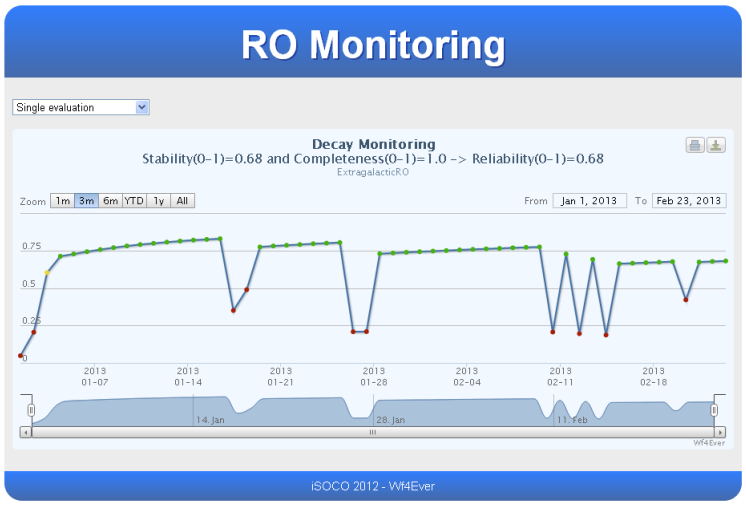

Fig. 2. RO-Monitor web application screenshot for the motivation scenario

Bob wants to reuse a workflow and because a research object contains much richer information for him to reuse the workflow, he starts with such a research object. The first step that he takes is to inspect the RO reliability trace for the RO of his interest. He can see at the beginning of the trace that the RO was initially created some time ago and afterwards its reliability increased due to the addition of some resources. Later on, he observes that there is a first drop on the reliability score, which was caused by a modification of one of the web services that was used by the workflow (i.e. the input format has changed for adopting ObsTAP VO standards). Once the input format is fixed by adopting the standard, the reliability increases; but it still needs more curation by modifying a script that was using the inputs that were changed previously. The second time the reliability drops is due to a period of time where the provider of web services and input data, which turns out to be Bob's institution, has stopped hosting them. When the provider restored the services, the RO reliability recovered and increased along the time until it suffered a successive set of problems related to the services, which were caused again by Bob's institution. This leads to a decrease in the reliability due to these workflow decay problems. The last reliability drop is caused by a script error when a data provider modified its output format from HTML to VOTable.

As we can see, by this reliability trace, Bob can obtain a much more complete picture of the changes of the workflow reliability over a time period, and more interestingly, an explanation behind these changes. This bigger picture as well as the explanations no doubt provide Bob with much more evidence for making decisions about the reliability of the workflow, and hence its reuse.

7 Conclusions and Future Work

Scientists, particularly computational scientists, are demanding new publication paradigms that pay more attention to the methods by which the results reported in publications were obtained. Amongst the various objectives of this movement, it is worthwhile highlighting some of the following, such as the need for validating the experiment, ensuring that the results are reproducible and therefore trustworthy as the basis of subsequent research, or, more generally speaking, making science more robust, transparent, pragmatic, and useful.

The work presented in this paper falls within such objectives. In particular it aims at contributing to the conservation and reuse of the published scientific methods, where reliability plays an important role. However, reliability cannot be drawn simply based on face value. We show evidence that, even in the case they were actually runnable and reproducible at the moment of publication, scientific workflows encoding such methods can experience decay due to different causes. When this happens, the reliability of the workflow, i.e. its claimed capability, could have been seriously undermined without careful consideration.

In this paper, we present our approach that is able to provide a more complete picture of changes that may occur to a workflow over a time period, to assist scientists to establish a more truthful indication of its reliability. Our results show evidence that the minimal set of information that we identified as necessary to be associated within a research object can indeed enable us to effectively assess some specific quality dimensions of a workflow at a time point and to monitor the change of this quality measure over a time period. Evidence is also shown that the completeness, stability and reliability metrics presented herein have the right behaviour to help scientists decide whether or not to reuse existing work for their own experiments and future publications.

We believe the new paradigm of semantic publications can benefit from our approach, supporting the incremental development and publication of new scientific advances based on the reuse of reproducible and reliable previous work. Our next steps will focus on the evaluation of the approach at a sufficiently large scale in specific communities of scientists in Astronomy and Biology. To this purpose, we are collaborating with scientific publishers like Gigascience interested in the application of our methods and tools in their publication pipeline.

Acknowledgments. The research reported in this paper is supported by the EU Wf4Ever project (270129) funded under EU FP7 (ICT-2009.4.1).

8 References

References

1. S. Bechhofer, I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch, D. Cruickshank, M. Delderfield, I. Dunlop, M. Gamble, D. Michaelides, S. Owen, D. Newman, S. Sufi, and C. Goble. Why linked data is not enough for scientists. Future Generation Computer Systems, 2011.

2. K. Belhajjame, O. Corcho, D. Garijo, J. Zhao, P. Missier, D. Newman, R. Palma, S. Bechhofer, E. García-Cuesta, J.M. Gómez-Pérez, G. Klyne, K. Page, M. Roos, J.E. Ruiz, S. Soiland-Reyes, L. Verdes-Montenegro, D. De Roure, and C.A. Goble. Workflow-centric research objects: First class citizens in scholarly discourse. In Proceeding of SePublica2012, pages 112, 2012.
3. C. Bizer. Quality-Driven Information Filtering in the Context of Web-Based Information Systems. VDM Verlag, 2007.
4. P.E. Bourne, T. Clark, R. Dale, A. De Waard, I. Herman, E. Hovy, D. Shotton, et al. Improving future research communication and escholarship: a summary of findings macquarie university researchonline. 2012. http://force11.org/white paper.
5. P. Ciccarese, M. Ocana, L.J. Garcia Castro, S. Das, and T. Clark. An open annotation ontology for science on web 3.0. J Biomed Semantics, 2(Suppl 2):S4, 2011.
6. D. De Roure, C. Goble, and R. Stevens. The design and realisation of the myexperiment virtual research environment for social sharing of workflows. Future Generation Computer Systems, 25:561567, 2009.
7. 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.
8. F. Giunchiglia and R. ChenuAbente. Scientific knowledge objects v. 1. Technical report, Technical Report DISI-09-006, University of Trento, 2009.
9. C.A. Goble, D. De Roure, and S. Bechhofer. Accelerating scientists knowledge turns. In Proceedings of The 3rd international IC3K joint conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 2012.
10. J.M. Gómez-Pérez, O. Corcho. Problem-Solving Methods for Understanding Process executions. Computing in Science and Engineering (CiSE), vol. 10, no. 3, pp. 47-52, May/June, 2008.
11. A. Mos, C. Pedrinaci, G. Alvaro, J.M. Gómez-Pérez, D. Liu, G. Vaudaux-Ruth, S. Quaireau, ServiceWave 2009, in Proceedings of the 2009 International Conference on Service-oriented Computing pp269-282.
12. P. Groth, A. Gibson, and J. Velterop. The anatomy of a nanopublication. Information Services and Use, 30(1):5156, 2010.
13. J. Hunter. Scientific publication packagesa selective approach to the communication and archival of scientific output. International Journal of Digital Curation, 1(1):3352, 2008.
14. P, Lord, S. Cockell, R. Stevens. Three Steps to Heaven: Semantic Publishing in a Real World Workow. In Proceeding of SePublica2012, pages 2334, 2012.
15. P. Mates, E. Santos, J. Freire, and C.T. Silva. Crowdlabs: Social analysis and visualization for the sciences. In SSDBM, pages 555564. Springer, 2011.
16. P. Missier, 2008. Modelling and computing the quality of information in e-science. Ph.D. thesis, School of Computer Science, University of Manchester.
17. 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.
18. Open archives initiative object reuse and exchange, 2008.
19. Page, K., Palma, R., Houbowicz, P., et. al. (2012). From workflows to Research Objects: an architecture for preserving the semantics of science. In Proceedings of the 2nd International Workshop on Linked Science.
20. J. Zhao, J.M. Gómez-Pérez, K. Belhajjame, G. Klyne, E. García-Cuesta, Garrido A, Hettne K, Roos M, De Roure D, Goble CA. Why Workflows Break - Understanding and Combating Decay in Taverna Workflows. In the proceedings of the IEEE eScience Conference (eScience 2012), IEEE CS, Chicago, USA, 2012.