The 13th International Semantic Web Conference
19-23 October 2014

LISC2014
Proceedings of the 4th Workshop on Linked Science
—
Making Sense Out of Data

Collocated with the 13th International Semantic Web Conference (ISWC2014)
Riva del Garda, Trentino, Italy

Editors:
Jun Zhao, Marieke van Erp, Carsten Keßler, Tomi Kauppinen,
Jacco van Ossenbruggen, Willem Robert van Hage

Preface

Traditionally scientific dissemination has been relying heavily on publications and presentations. The findings reported in these articles are often backed by large amounts of diverse data produced by complex experiments, computer simulations, and observations of physical phenomena. Although publications, methods and datasets are often related, due to this avalanche of data it remains extremely hard to correlate, reuse and leverage scientific data. Semantic Web technologies provide a promising means for publishing, sharing, and interlinking data to facilitate data reuse and the necessary correlation, integration, and synthesis of data across levels of theory, techniques and disciplines. However, even when these data become discoverable and accessible, significant challenges remain in making intelligent understandings of these data and scientific discoveries that we anticipated.

Our past three series (LISC2011, LISC2012 and LISC2013) have seen many novel ideas of using Semantic Web technologies for integrating scientific data (for example about real experiments or from simulations), or enabling reproducibility of research via online tools and Linked Data. The theme for LISC2014 is "Making Sense out of Data Through Linked Science". Here we focus on new ways of discovering interesting patterns from scientific data, which could lead to research validation or identification of new hypotheses and acceleration of the scientific research cycle. We target both new results through making use of semantic reasoning or making innovative combination of existing technologies (such as visualization, data mining, machine learning, and natural language processing) with SW technologies to enable better understanding of data. One goal is to create both an incentive for scientists to consider the Linked Science approach for their scientific data management and an incentive for technologists from different disciplines to work together towards the vision of powering science with technologies.

LISC2014 was hosted at the 13th International Semantic Web Conference (ISWC2014), in Riva del Garda, Trentino, Italy. Twenty-seven attendees enjoyed the opening keynote "Making more sense out of social data" by Harith Alani (KMI, the Open University, UK), followed by excellent presentations of the eight regular papers collected in these proceedings. We continued the tradition of a "working" workshop with a plenary discussion on the challenges and opportunities of using Semantic Web technologies for sense making. The results of this discussion is published at FigShare, and can be cited as:

Zhao, Jun; Patton, Evan; Vardeman, Charles; Peroni, Silvio; Osborne, Francesco; Nart, Dario De; Dumontier, Michel; Diallo, Gayo; van Ossenbruggen, Jacco (2014):
LISC 2014 - Results: Discussion on Challenges in Making Sense Out Of Data Using Linked Data Technologies.
`http://dx.doi.org/10.6084/m9`. figshare. 1209243

Overall, this edition continued providing a successful forum for discussing how semantic web technologies and linked data can help science. We wanted to thank the entire program committee for helping to assemble the program and the attendees for their enthusiastic participation.
The LISC 2014 Co-organizers:

Jun Zhao
Marieke van Erp
Carsten Keßler
Tomi Kauppinen
Jacco van Ossenbruggen
Willem Robert van Hage

Program Committee

Boyan Brodaric
Arne Broering
Paolo Ciccarese
Oscar Corcho
Aba-Sah Dadzie
Stefan Dietze
Mathieu Daquin
Daniel Garijo
Alasdair Gray
Paul Groth
Rinke Hoekstra
Krzysztof Janowicz
Simon Jupp
Tomi Kauppinen
Carsten Keßler
James Malone
Edgard Marx (additional reviewer)
Jeff Pan
Heiko Paulheim
Marieke van Erp
Willem van Hage
Jacco van Ossenbruggen
Amrapali Zaveri
Jun Zhao

Contents

EPUB3 for Integrated and Customizable Representation of a Scientific Publication and its Associated Resources

Hajar Ghaem Sigarchian[1], Ben De Meester[1], Tom De Nies[1], Ruben Verborgh[1], Wesley De Neve[1,2], Erik Mannens[1], and Rik Van de Walle[1]

[1] Ghent University - iMinds - Multimedia Lab
Gaston Crommenlaan 8 bus 201, B-9050 Ledeberg-Ghent, Belgium
[2] Korea Advanced Institute of Science and Technology (KAIST) - IVY Lab
Yuseong-gu, Daejeon, Republic of Korea
{hajar.ghaemsigarchian, ben.demeester, tom.denies, ruben.verborgh, wesley.deneve, erik.mannens, rik.vandewalle}@ugent.be

Abstract. Scientific publications point to many associated resources, including videos, prototypes, slides, and datasets. However, discovering and accessing these resources is not always straightforward: links could be broken, readers may be offline, or the number of associated resources might make it difficult to keep track of the viewing order. In this paper, we explore potential integration of such resources into the digital version of a scientific publication. Specifically, we evaluate the most common scientific publication formats in terms of their capability to implement the desirable attributes of an enhanced publication and to meet the functional goals of an enhanced publication information system: PDF, HTML, EPUB2, and EPUB3. In addition, we present an EPUB3 version of an exemplary publication in the field of computer science, integrating and interlinking an explanatory video and an interactive prototype. Finally, we introduce a demonstrator that is capable of outputting customized scientific publications in EPUB3. By making use of EPUB3 to create an integrated and customizable representation of a scientific publication and its associated resources, we believe that we are able to augment the reading experience of scholarly publications, and thus the effectiveness of scientific communication.

1 Introduction

Scientific publications consist of more than only text: they may also point to many associated (binary) resources, including videos, prototypes, slides, and datasets. Yet today, only the access to the *text* of a scientific publication is straightforward; the associated resources are often more difficult to access. For instance, readers may not always have an Internet connection at their disposal to download related materials, and even when this is the case, links might become broken after a while. Furthermore, given their diverse nature, related materials often need to be accessed in a different reading environment like a standalone media player, causing readers to lose track of the scientific narrative.

The 2007 Brussels Declaration[3] by the International Association of Scientific, Technical and Medical (STM) Publishers states that "raw research data should be made freely available" and that "one size fits all solutions will not work". In this paper, we illustrate that the ability to (adaptively) create an integrated representation of a scientific publication and its associated resources contributes to these goals. Specifically, we evaluate the most common scientific publication formats in terms of their capability to implement the desirable attributes of an enhanced publication and to meet the functional goals of an enhanced publication information system: PDF, HTML, EPUB2, and EPUB3. In addition, we present an EPUB3 version of an exemplary publication in the field of computer science, integrating and interlinking an explanatory video and an interactive prototype. Finally, we introduce a demonstrator that is capable of outputting customized scientific publications in EPUB3.

The rest of this paper is structured as follows. In Section 2, we discuss a number of current best practices among three scientific publishers, focusing on the way open formats and their features are used to enhance scientific publications. Next, in Section 3, we investigate to what extent PDF, HTML, EPUB2, and EPUB3 facilitate the use of enhanced scientific publications and corresponding information systems. In Section 4, we present an exemplary scientific publication in EPUB3 that integrates an explanatory video and an interactive prototype. In Section 5, we introduce our demonstrator for creating customized scientific publications in EPUB3. Finally, in Section 6, we present our conclusions and a number of directions for future work.

2 Current Best Practices

In this section, we briefly discuss a number of current best practices among three scientific publishers, focusing on the way open formats are used to make available scientific publications that have been enhanced with multimedia, interactivity, and/or Semantic Web features.

BioMed Central and Hindawi Publishing Corporation: These publishers make scientific publications available in several formats: PDF, HTML, and EPUB2. The HTML version of the publications can for instance be enhanced with reusable data (e.g., supplementary datasets), while the EPUB2 version of the publications just uses links to cited publications in EPUB2 format. However, the publications in question do not contain any embedded interactive multimedia content.

Elsevier: Elsevier makes available different versions of a scientific publication: PDF, HTML, MOBI, and EPUB2. In addition, authors are able to deposit their datasets, making it possible for readers to access and download these datasets [1]. Moreover, the EPUB2 version of a publication is enriched with direct links to the PDF version of cited publications, thus not embedding these PDF versions into the EPUB2 file. Furthermore, the EPUB2 version of a publication does not contain any embedded interactive multimedia content.

[3] `http://www.stm-assoc.org/brussels-declaration/`

In summary, we can conclude that none of the aforementioned EPUB2 versions – as currently made available by BioMed Central, Hindawi Publishing Corporation, and Elsevier – embed interactive multimedia content for offline usage (i.e., readers need to have network connectivity in order to be able to access all linked resources), nor do they contain Semantic Web features.

3 Comparative Analysis of Publication Formats

In recent years, a new open format for distribution and interchange of digital publications has emerged, called EPUB3 [6]. This format can also be used in the context of scientific publications. In what follows, we investigate to what extent PDF, HTML, EPUB2, and EPUB3 are able to support the properties of an enhanced scientific publication (that is, a scientific publication with multimedia, interactivity, and/or Semantic Web features). To that end, we analyzed a number of desirable attributes of an enhanced publication. Furthermore, we also investigated the functional goals of an enhanced publication information system (that is, the system that facilitates the authoring of enhanced publications).

Thoma *et al.* [10] defined a core set of nine desirable attributes of an enhanced publication: *appearance, page transitions, in-page navigation, image browsing, navigation to an embedded/linked media object, support for interactivity, transmission, embedding and linking of multimedia/interactive objects,* and *document integrity and structure*. In addition, by both considering the attributes defined by Thoma *et al.* in [10] and a review of five already existing enhanced publications, Adriaansen *et al.* [2] identified eleven attributes of an enhanced publication: *navigation by table of contents, metadata, links to figures and tables, attached data resources, link from text to references, direct publication links from references, reader comments, download as PDF, interactive content, relations,* and *cited by*. Furthermore, as argued in a talk by Ivan Herman[4], *bridging online and offline access* is a need for high-quality digital books, and consequently for high-quality digital scientific publications, given that offline access enables users to access supplementary information, even when they do not have a network connection at their disposal. As a result, although none of the aforementioned research efforts discusses this aspect, we consider offline access to be a desirable attribute of an enhanced publication as well.

Besides the attributes of enhanced publications, we also considered data model and information system aspects. Bardi *et al.* [3] reviewed existing data models for enhanced publications, taking into account structural and semantic features, also proposing a classification scheme for enhanced publication information systems based on their main functional goals. In this context, the authors outline four major scientific motivations that explain the functional goals of an enhanced publication information system: *packaging with supplementary material, improving readability and understanding, interlinking with research data*, and *enabling repetition of experiments*. Furthermore, we believe that *portability*

[4] http://www.w3.org/2014/Talks/0411-Seoul-IH/Talk.pdf

is also needed in order to preserve the availability of resources and their interlinking, given that it enables users to even access supplementary information in offline situations. Thus, an enhanced publication that has supplementary resources needs to be a self-contained package. Therefore, we identified *portable packaged file* as another desirable attribute of an enhanced publication.

Finally, according to Liu [8], users are in need of a hybrid solution for print and digital resources. This means that, besides all different digital publication formats, *print* also remains an important publication medium. As a result, we see *suitable for print* as another desirable attribute of an enhanced publication.

Ideally, an enhanced publication information system should be able to support all the desirable attributes mentioned above. Considering the desirable attributes of enhanced publications and the functional goals of enhanced publication information systems, we mapped the attributes identified in [10,2] onto each functional goal identified by Bardi *et al.* in [3]. Our mapping can be found in the first and second column of Table 1. We can observe that nearly all desirable attributes of an enhanced publication can be covered by the functional goals of an enhanced publication information system, with the exception of the final three attributes, for which we defined our own functional goals.

Next, we investigated what scientific publication formats are the most promising to cover both the desirable attributes of an enhanced publication and the functional goals of an enhanced publication information system. We have summarized our findings in the four rightmost columns of Table 1. Corresponding explanatory notes can be found below.

Packaging with supplementary material: This functional goal states that it should be possible to add supplementary material to a scientific publication. PDF can embed audio and video but it does not support rich media (e.g., media overlays). As such, it is not a suitable format for embedding various types of associated resources (e.g., interactive content and standalone applications). Consequently, PDF has limited support for this functional goal and its underlying attributes. Note that extensions exist, such as export to a PDF Portfolio in Adobe Acrobat[5], that make it possible to combine related materials. However, to the best of our knowledge, none of these extensions for instance allow embedding interactive content and standalone applications. Furthermore, the embedded resources are not reusable, unlike the EPUB3 format, which lets users reuse embedded resources. In order to package research data within an HTML file, all the dependencies need to be packaged as well. While this is possible (e.g., using a zipped folder), there is no standardized approach to do this, as opposed to EPUB2 and EPUB3. Therefore, we do not consider HTML to be suitable for meeting this functional goal. According to the EPUB2 specification [7], EPUB2 cannot embed multimedia and interactive objects. Consequently, EPUB2 also offers limited support for this functional goal. However, in EPUB3, no such restrictions are specified. As a result, we can conclude that EPUB3 is the only format that fully supports this functional goal.

[5] `http://www.adobe.com/products/acrobat/combine-pdf-files-portfolio.html`

Functional Goals	Attributes	Format			
		PDF	HTML	EPUB2	EPUB3
Packaging with supplementary material	– Embedding and linking of multimedia/interactive objects – Document integrity and structure – Attached data resources – Navigating to an embedded / linked media object	✓*			✓
Enabling repetition of experiments	– Native support for interactivity – Code execution – Interactive content		✓		✓
Improving readability and understanding	– Navigation by table of contents – Reader comments – Appearance – Page transitions – In-page navigation – Image browsing – Links to figures and tables – Direct publication links from references – Cited by	✓*	✓*	✓	✓
Interlinking with research data	– Metadata – Relations		✓	✓*	✓
Portable packaged file	– Bridging online / offline – Transmission	✓*		✓*	✓
Suitable for print	– Download as PDF	✓			

Table 1: Support for enhanced publication attributes (* = limited support).

Enabling repetition of experiments: This functional goal aims at enabling researchers to (re-)execute experiments and/or demonstrators from within a scientific publication. PDF has limited support for scripting and code execution. However, the support available is not sufficient for building small standalone applications that can act as interactive content (e.g., self-contained widgets). As a result, PDF is not suitable for meeting this functional goal. HTML is able to embed code (e.g., JavaScript). Moreover, thanks to the inline frame element (that is, the `iframe` element), HTML can also be used as an interface to other experiments. As EPUB2 does not support JavaScript, it is not suited for repetition of experiments. However, similar to HTML, EPUB3 supports JavaScript, and thus the aforementioned functional goal

(unless experiments are involved that for instance use complex algorithms on clusters to obtain their results).

Improving readability and understanding: PDF is a specific format for print, and not for screen readers. While still undeniably the most suitable format for print layout, in digital form, it does not have device independence [5], making it difficult to maintain readability on different screens. According to the PDF specification, it has a limited support for this functional goal. On the other hand, HTML, EPUB2, and EPUB3 are suitable for improving readability and understanding, because they can overcome the aforementioned shortcomings of PDF (*cf.* the use of reflowable layout).

Interlinking with research data: In order to make links between supplementary materials added to publications, (relational) metadata need to be taken into account. PDF has a coarse level of support for metadata (e.g., title and author information), and where these metadata are not related to interlinking supplementary materials. As a result, PDF is not suitable for meeting this functional goal. HTML can be enriched for interlinking purposes using Semantic Web formats and technologies [9] (e.g., RDF and OWL). EPUB2 has limited support for metadata. Furthermore, it does not allow embedding multimedia and interactive content as supplementary research data. Hence, EPUB2 is not suitable for meeting this functional goal. According to the EPUB3 specification, it supports metadata and interlinking of research data. In fact, it retains all functionality of (X)HTML5.

Apart from a suitable format, interlinking supplementary materials requires suitable ontologies. Fortunately, many suitable candidates for general and specific interlinking purposes are already available. For example, `schema.org` is an ontology that is suitable for use in a variety of domains, including the description of events and creative works. It can thus be used to semantically enhance publications, and it can also be extended by other ontologies. Furthermore, Standard Analytics[6] aims at turning scholarly publications into an interface to a web of data, making use of already existing web ontologies. Moreover, Structural, Descriptive, and Referential (SDR)[7] is an ontology for representing academic publications, related artifacts (e.g., videos, slides, and datasets), and referential metadata. This ontology can generically define all possible interactive and multimedia resources. In addition, any publication can use general ontologies such as the Citation Typing Ontology (CiTO)[8], the Bibliographic Ontology (BIBO)[9], and the Common European Research Information Format (CERIF)[10]. Finally, publications may also need to make use of ontologies that are specific for their research domains (e.g., in the medical domain, the Infectious Disease Ontology (IDO)[11] could be used).

[6] `https://standardanalytics.io/`
[7] `http://onlinelibrary.wiley.com/doi/10.1002/asi.23007/full`
[8] `http://www.essepuntato.it/lode/http://purl.org/spar/cito`
[9] `http://bibliontology.com/`
[10] `http://helios-eie.ekt.gr/EIE/bitstream/10442/13864/1/IJMSO_2014_CERIF_authorFinalVersion.pdf`
[11] `http://infectiousdiseaseontology.org/page/Main_Page`

Portable packaged file: PDF has limited support for packaging interactive content and standalone applications. Furthermore, it cannot bridge the gap between online and offline usage. Indeed, PDF is an offline format for print, and any interactive parts will not remain after printing a publication. As mentioned before, HTML lacks a proper packaging structure, making this format not a suitable candidate for meeting this functional goal. A similar remark holds regarding EPUB2, as this format does not have support for embedding interactive multimedia resources. As EPUB3 has extensive support for embedding interactive multimedia resources, it can be considered a suitable format for creating portable packaged files. Ideally, users expect that all types of resources can be embedded in a packaged file, regardless of their size. This is one of the shortcomings of EPUB3. Embedding large datasets makes the size of an EPUB3 file potentially very large, causing portability and readability issues. We discuss a possible solution to this issue in Section 5.

Suitable for print: Currently, PDF is the only format suitable for print. Although HTML, EPUB2, and EPUB3 can also be used for the purpose of print, they have been designed for screen readers and can currently not match the high typesetting demands for print publications.

As can be seen in Table 1, EPUB3 is the format that supports most desirable attributes of an enhanced publication and most functional goals of an enhanced publication information system. Only PDF is suitable for print output, given that HTML and EPUB(2/3) have been primarily designed for screen output, typically resulting in a layout that is suboptimal for print. Note that, as a workaround for this problem, the EPUB(2/3) and HTML versions of a publication can embed or link to the PDF version of a publication.

4 Proof-of-Concept: A Scientific Publication in EPUB3

In this section, we demonstrate how EPUB3 can be used to create an integrated representation of a scientific publication and its associated resources. To that end, we enhanced the "Everything is Connected" publication [11] – a paper authored by ourselves and a number of colleagues – embedding an explanatory video and an interactive prototype. The resulting proof-of-concept is available for download[12]. We used Readium[13] as our electronic reading system, since it supports most features of EPUB3. As illustrated by Figure 1, our proof-of-concept shows how a publication can act as an interface to different types of research outputs. Note that, instead of adding a link to the online version of the interactive prototype, we made use of an `iframe` to allow immediate access to the interactive prototype from within the publication, thus not requiring the reader to make use of a different reading environment.

[12] `http://multimedialab.elis.ugent.be/users/hghaemsi/EnhancedPublication.epub`

[13] `http://readium.org/`

Furthermore, we semantically enhanced our exemplary EPUB3 publication by making use of `schema.org`, a general ontology that allows describing books and articles, among other creative works. Thanks to properties such as `embedUrl`, `description`, and `contentUrl`, `schema.org` makes it possible to indicate how a resource is related to the target EPUB3 publication in a straightforward way. We illustrate this in Figure 2. Note that `schema.org` is supported by major search engines such as Bing, Google, Yahoo!, and Yandex. However, at the time of writing this paper, the aforementioned search engines did not have support yet for indexing EPUB3 publications (and reading the metadata available within these publications).

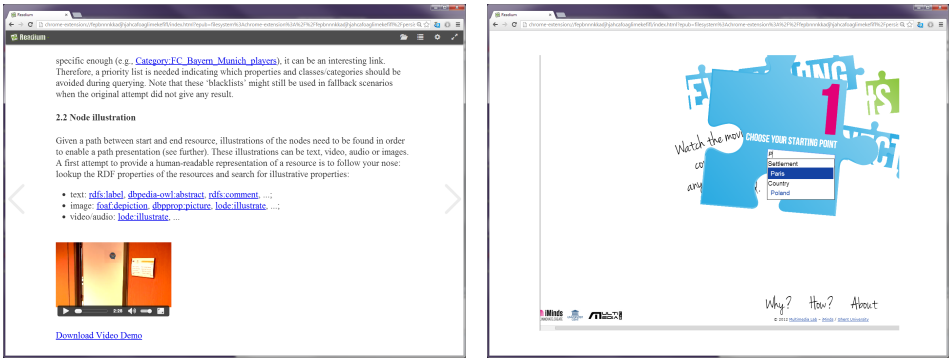


Fig. 1: Exemplary scientific publication enhanced with an explanatory video (left) and an interactive prototype (right). Both the video and the prototype have been embedded into the EPUB3 version of the scientific publication.

```
<div vocab="http://schema.org/" property="video" typeof="VideoObject">
  <p>Below, you can find <span  property="Description">an embedded screencast</span>.</p>
  <video width="320" height="240" controls="">
    <source property="embedUrl" src="./video.mp4" type="video/mp4"/>
  </video>
  <p>This screencast can also be accessed <a property="url" href="http://youtu.be/FawygFT5Brs">remotely</a>,
    or can be <a property="contentUrl" href="./video.mp4">downloaded</a>.</p>
</div>
```

Fig. 2: Use of `schema.org` for interlinking a local and remote video object.

5 Creating Customized EPUB3 Publications

In the previous sections, we explained how supplementary materials can be embedded into a scientific publication. As mentioned before, embedding all relevant

supplementary materials in a portable packaged file is not always cost-effective and/or desirable for a user. Since the size of an EPUB3 file is dependent on the size of all embedded resources, it will not be lightweight in all use cases, e.g., when embedding large datasets. The problem is that, on the one hand, a packaged file should not face portability and other usage issues relevant to its size. On the other hand, the advantages of having a portable packaged publication are overthrown with the disadvantage of not being able to distribute the entire publication properly. Users may not need all embedded supplementary materials and instead, wish to have their own customized lightweight publication. For instance, we can refer to big datasets or high-resolution images which can be located in a remote repository instead of embedding them in the portable packaged file. An environment for outputting customized publications allows users to select and embed the supplementary materials to the extent that they choose. Hence, they can determine the size of the EPUB3 file themselves. That way, the problem of distributing overly large publications is solved, and only the content that the user needs is distributed. The only disadvantage of this approach is the added complexity at the distribution side (i.e., at the platform of the publisher). However, most publishers already have an extensive online distribution infrastructure, which could easily be expanded with an interface such as the one we propose. For example, publishers such as Elsevier offer different formats of a publication to users. In particular, on the ScienceDirect website of Elsevier, there is an option for the user to select his/her preferred format.

To illustrate this concept of customizable publications, we implemented a basic demonstrator in which a user can first select the relevant supplementary material using a web interface, after which a customized EPUB3 publication is outputted. Figure 3 shows the user interface of our online demonstrator. Content selection is entirely done at the client side, based on the HTML representation of a publication. The selected content is then packaged as an EPUB3 file on the server side. The resulting demonstrator is available online[14]. Note that the author of a publication can determine which elements are customizable, simply by adding the class `customizable` to the desired HTML elements.

Ideally, the implemented functionality for outputting customized publications in EPUB3 would be integrated into an authoring environment, where authors and publishers could indicate which elements of a publication are customizable. In previous work, we have implemented such an authoring environment for the collaborative creation of enriched e-Books using EPUB3 [4]. It allows authors and publishers to create an electronic publication with all required material embedded. Next, this publication can be exported as an EPUB3 file. In future work, we aim to showcase an integrated version of this authoring environment with a customizable distribution platform as described above.

[14] `http://uvdt.test.iminds.be/custompublication/books/1/main.xhtml`

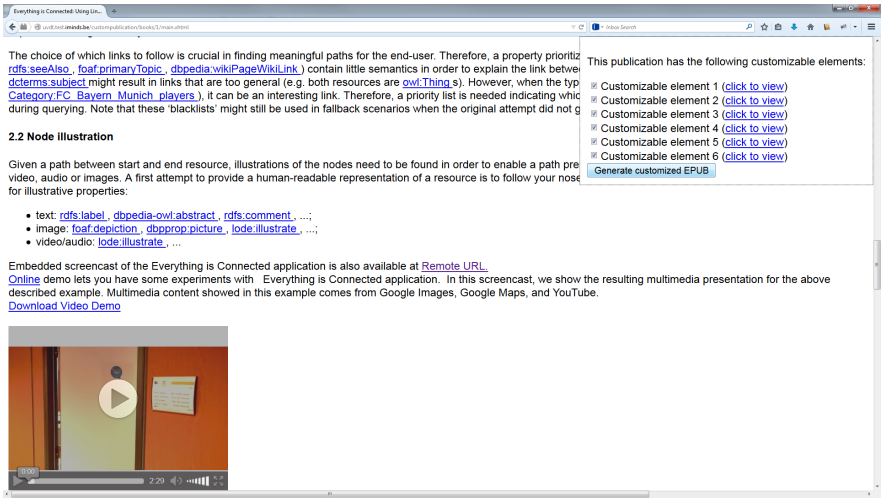


Fig. 3: The interface of our demonstrator for creating customized publications. Users can select the supplementary materials that they want to have embedded in the EPUB3 version of the enhanced publication.

6 Conclusions and Future Work

In this paper, we demonstrated that the increasingly popular EPUB3 format can be used to create integrated representations of a scientific publication and its associated resources. By doing so, we believe that this contributes to a better reading experience and more effective scientific communication (e.g., support for the inclusion of explanatory videos and interactive prototypes should enable authors to better transfer their knowledge and experience). In addition, we indicated that an EPUB3 version of a scientific publication can be used as a primary version, from which other versions of the scientific publication can be reached (e.g., a PDF version for print), thereby allowing legacy content to persist.

We can identify a number of directions for future research. First, user-friendly authoring tools are needed that allow easily creating enhanced scientific publications, and where these scientific publications can act as an interface to different research outputs. We have already started taking steps in this direction. Second, these authoring tools need to support different output formats, in order to meet the needs of both readers that are reading on paper and readers that are reading digitally. Third, these authoring tools also need to make it possible to easily add metadata to EPUB3 versions of scientific publications, such that EPUB3 versions of scientific papers may have the same degree of discoverability as PDF and HTML versions. Finally, it would be interesting to investigate the good practices of novel publication repositories such as PLOS ONE, Figshare, and ResearchGate.

7 Acknowledgments

The research activities described in this paper were funded by Ghent University, iMinds (a research institute founded by the Flemish Government), the Institute for Promotion of Innovation by Science and Technology in Flanders (IWT), the FWO-Flanders, and the European Union.

References

1. Aalbersberg, I.J., Dunham, J., Koers, H.: Connecting Scientific Articles with Research Data: New Directions in Online Scholarly Publishing. Data Science Journal 12(0), WDS235–WDS242 (2013)
2. Adriaansen, D., Hooft, J.: Properties of Enhanced Publications and the Supporting Tools
3. Bardi, A., Manghi, P.: Enhanced Publications: Data Models and Information Systems. LIBER Quarterly 22 (2014)
4. De Meester, B., De Nies, T., Ghaem Sigarchian, H., Vander Sande, M., Van Campen, J., Van Impe, B., De Neve, W., Mannens, E., Van de Walle, R.: A Digital-First Authoring Environment for Enriched e-Books using EPUB 3. In: Proceedings of the 18th Int'l. Conference on Electronic Publishing (ELPUB), June 19-20, Thessaloniki, Greece (2014)
5. Eikebrokk, T., Dahl, T.A., Kessel, S.: EPUB as Publication Format in Open Access Journals: Tools and Workflow. Code4Lib Journal 24 (2014)
6. IDPF: Electronic Publication, version 3. `http://idpf.org/epub/30`
7. IDPF: Open Publication Structure (OPS). `http://www.idpf.org/epub/20/spec/OPS_2.0.1_draft.htm#Section1.3.7`
8. Liu, Z.: Print vs. Electronic Resources: A Study of User Perceptions, Preferences, and Use. Information Processing & Management 42(2), 583–592 (2006)
9. Shotton, D.: Semantic Publishing: The Coming Revolution in Scientific Journal Publishing. Learned Publishing 22(2), 85–94 (2009)
10. Thoma, G.R., Ford, G., Chung, M., Vasudevan, K., Antani, S.: Interactive Publications: Creation and Usage. In: Electronic Imaging 2006. pp. 607603–607603. International Society for Optics and Photonics (2006)
11. Vander Sande, M., Verborgh, R., Coppens, S., De Nies, T., Debevere, P., De Vocht, L., De Potter, P., Van Deursen, D., Mannens, E., Van de Walle, R.: Everything is Connected. In: Proceedings of the 11th International Semantic Web Conference (ISWC) (2012)

Semantic lenses to bring digital and semantic publishing together

Angelo Di Iorio[1], Silvio Peroni[1,2], Fabio Vitali[1], and Jacopo Zingoni[1]

[1] Department of Computer Science and Engineering, University of Bologna (Italy)
{angelo.diiorio, silvio.peroni, fabio.vitali, jacopo.zingoni}@unibo.it
[2] STLab-ISTC, Consiglio Nazionale delle Ricerche (Italy)

Abstract. Modern scholarly publishers are making steps towards *semantic publishing*, i.e. the use of Web and Semantic Web technologies to represent formally the meaning of a published document by specifying information about it as metadata and to publish them as Open Linked Data. In this paper we introduced a way to use a particular semantic publishing model, called *semantic lenses*, to semantically enhance a published journal article. In addition, we present the main features of *TAL*, a prototypical application that enables the navigation and understanding of a scholarly document through these semantic lenses, and we describe the outcomes of a user testing session that demonstrates the efficacy of TAL when addressing tasks requiring deeper understanding and fact-finding on the content of the document.

Keywords: Web interface, document semantics, semantic publishing

1 Introduction

Simultaneously to the evolution of the Web by means of Semantic Web technologies, modern publishers (and in particular scholarly publishers) are making steps towards the enhancing of digital publications with semantics, an approach that is known as *semantic publishing* [22]. In brief, semantic publishing is the use of Web and Semantic Web technologies to represent formally the meaning of a published document by specifying a large quantity of information about it as metadata and to publish them as Open Linked Data. As a confirmation of this trend, recently the Nature Publishing Group (publisher of *Nature*), the American Association for the Advancement of Science (publisher of *Science*) and the Oxford University Press have all announced initiatives to open their articles' reference lists and to publish them as Open Linked Data[3,4,5].

[3] Nature.com Linked Data: http://data.nature.com.
[4] http://opencitations.wordpress.com/2012/06/16/science-joins-nature-in-opening-reference-citations
[5] http://opencitations.wordpress.com/2012/06/22/oxford-university-press-to-support-open-citations

However, the enhancement of a traditional scientific paper with semantic annotations is not a straightforward operation, since it involves much more than simply making semantically precise statements about named entities within the text. In [17], we have shown how several relevant points of view exist beyond the bare words of a scientific paper – such as the context of the publication, its structural components, its rhetorical structures (e.g. Introduction, Results, Discussion), or the network of citations that connects the publication to its wider context of scholarly works. These points of view are usually combined together to create an effective unit of scholarly communication so well integrated into the paper as a whole and into the rhetorical flow of the natural language of the text, so as to be scarcely discernible as separate entities by the reader. We also propose the separation of these aspects into eight different sets of machine-readable semantic assertions (called *semantic lenses*), where each set describes one of (from the most contextual to the most document-specific): research context, authors' contributions and roles, publication context, document structure, rhetoric organisation of discourse, citation network, argumentative characterisation of text, and textual semantics.

How can the theory of semantic lenses be used to extend effectively semantic publishing capabilities of publishers? In order to provide an answer to this question, in this paper we introduce a prototypical HTML interface to scholarly papers called *TAL* (*Through A Lens*), which enables the navigation of a text document on which semantic lenses have been applied to make explicit all the corresponding information. This HTML interface is meant to be a proof of concept of the semantic lenses in a real-case scenario. We performed a user testing session that demonstrates the efficacy of TAL when addressing tasks requiring deeper understanding and fact-finding on the content of the document.

The rest of the paper is organised as follows. In Section 2 we introduce some significant works related to semantic publishing experiences and models. In Section 3 we show an application of semantic lenses onto a particular scholarly article. In Section 4 we introduce TAL describing its main features, while in Section 5 we discuss the outcomes of a user testing session we performed to assess the usability and effectiveness of TAL. Finally (Section 6) we conclude the paper sketching out some future works.

2 Related works

Much current literature concerns both the proofs of concepts for semantic publishing applications and the models for the description of digital publishing from different perspective. Because of this richness, here we present just some of the most important and significant works on these topics.

In [22], Shotton *et al.* describe their experience in enriching and providing appropriate Web interfaces for scholarly papers enhanced with provenance informations, scientific data, bibliographic references, interactive maps and tables, with the intention to highlights the advantages of semantic publishing to a broader audience. Along the same lines, in their work [19] Pettifer *et al.* introduce pros

and cons of the various formats for the publication of scholarly articles and propose an application for the semantic enhancement of PDF documents according to established ontologies.

A number of vocabularies for the description of research projects and related entities have been developed, e.g. the *VIVO Ontology*[6] – developed for describing the social networks of academics, their research and teaching activities, their expertise, and their relationships to information resources –, the *Description Of A Project*[7] – an ontology with multi-lingual definitions that contains terms specific for software development projects – and the *Research Object* suite of ontologies [1] – for linking together scientific workflows, the provenance of their executions, interconnections between workflows and related resources (datasets, publications, etc.), and social aspects related to such scientific experiments.

One of the most widely used ontology for describing bibliographic entities and their aggregations is BIBO, the *Bibliographic Ontology* [3]. FRBR, *Functional Requirements for Bibliographic Records* [10], is yet another more structured model for describing documents and their evolution in time. One of the most important aspects of FRBR is the fact that it is not tied to a particular metadata schema or implementation.

Several works have been proposed in the past to model the rhetoric and argumentation of papers. For instance, the SALT application [9] permits someone such as the author "to enrich the document with formal descriptions of claims, supports and rhetorical relation as part of their writing process". There are other works, based on [23], that offer an application of Toulmin's model within specific scholarly domains, for instance the legal and legislative domain [11]. A good review of all the others Semantic Web models for the description of arguments can be found in [21].

3 The Semantic Lenses

In [17], we claimed that the semantics of a document is definable from different perspectives, where each perspective is represented as a *semantic lens* that is *applied* to a document to reveal a particular semantic facet. In this section we briefly summarise our theory. A full example of the lenses applied to a well-known paper *Ontologies are us: A unified model of social networks and semantics* [14] is available at `http://www.essepuntato.it/lisc2014/lens-example`.

Lenses are formalised in the LAO ontology[8]. In addition, since the application of the semantic lenses to a document is an *authorial activity*, i.e. the action of a person (the original author as well as anyone else) taking responsibility for a semantic interpretation of the document, we also record the provenance of the semantic statements according to the *Provenance Ontology* (*PROV-O*) [12].

Figure 1 summarises the overall conceptual framework. The lenses are organised in two groups: *context*-related, which describe the elements contributing to

[6] VIVO Ontology: http://vivoweb.org/ontology/core
[7] DOAP: http://usefulinc.com/ns/doap
[8] Lens Application Ontology (LAO): http://www.essepuntato.it/2011/03/lens.

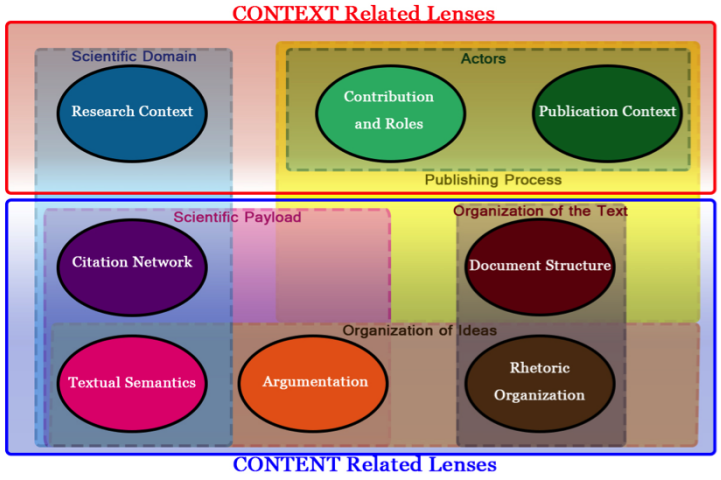


Fig. 1. The layout of Semantic Lenses in relation to the facets of a scientific document.

the creation and development of a paper, and *content*-related, which describe the content itself of the paper from different angles.

3.1 Describing the context

Writing a scientific paper is usually the final stage of an often complex collaborative and multi-domain activity of undertaking the research investigation from which the paper arises. The organizations involved, the people affiliated to these organizations and their roles and contributions, the grants provided by funding agencies, the research projects funded by such grants, the social context in which a scientific paper is written, the venue within which a paper appears: all these provide the research *context* that leads, directly or indirectly, to the genesis of the paper, and awareness of these may have a strong impact on the credibility and authoritativeness of its scientific content.

Three lenses are designed to cover these aspects:

- *Research context*: the background from which the paper emerged (the research described, the institutions involved, the sources of funding, etc.). To describe such *contextual environment* we use *FRAPO*, the *Funding, Research Administration and Projects Ontology*[9].
- *Contributions and roles*: the individuals claiming authorship on the paper and what specific contributions each made. We use *SCoRO* (the *Scholarly Contributions and Roles Ontology*[10]) and its imported ontology *PRO* (the *Publishing Roles Ontology*[11]) [18] to describe these aspects.

[9] FRAPO: http://purl.org/cerif/frapo
[10] SCoRO: http://purl.org/spar/scoro
[11] PRO: http://purl.org/spar/pro

– *Publication context*: any information about the event (e.g., a conference) and publication venue of the paper (such as the proceedings or the journal), as well as connections to the other papers sharing the same event or venue. This part is described by using FaBiO, the *FRBR-aligned Bibliographic Ontology*[12] [16] and BiRO, the *Bibliographic Reference Ontology*[13][5].

Note that all the ontologies used or suggested in this paper to describe "lenses" statements have been chosen as an appropriate and convincing example of an ontology that fulfils the requirements for the lens, since they allow us to fully describe all the document aspects we are interested in. However, their use is not mandatory, so as to leave people to use other models (such as those described in Section 2) instead of them.

3.2 Describing the content

The semantics of *the content* of a document, i.e. such a semantics that is implicitly defined in and inferable from the text, can be described from different points of view. For example, the semantical *structure* of the text – i.e. the organization of the document as structured containers, blocks of text, inline elements – is often expressed by means of markup languages such as XML and LaTeX, that have constructs for describing content hierarchically.

In a Semantic Web context, we would rather use an ontology that describes the markup structures in OWL. For this reason, we use *EARMARK* [8], an ontology[14] of a markup metalanguage, to describe the structure of the document as a set of OWL assertions to associate formal and explicit semantics [15]. Through the *Pattern Ontology* (*PO*)[15] [6] in combination with EARMARK we can associate a particular structural semantics to markup elements, such an element *h1 expressing* the concept of being a block of text, or the *div* element containing it being a container. This is covered by the *document structure* lens.

Close to that, we place the identification and organization of the *rhetorical components* of the text, such as a section being an *Introduction*, some paragraphs describing the *Methods* of the research, or the presented *Results* or the paper's *Conclusion*), in order to label all the meaningful aspects of the scientific discourse. Such rhetoric characterization of markup structures can be specified through *DoCO*, the *Document Components Ontology*[16], and *DEO*, the *Discourse Elements Ontology*[17].

In addition, strictly correlated with the rhetorical aspects of a document, we can detail the organization of the claims and the arguments of the paper (providing evidences to a claim) . The argumentative organisation of discourse is

[12] FaBiO: http://purl.org/spar/fabio
[13] BiRO: http://purl.org/spar/biro
[14] EARMARK: http://www.essepuntato.it/2008/12/earmark
[15] PO: http://www.essepuntato.it/2008/12/pattern
[16] DoCO: http://purl.org/spar/doco
[17] DEO: http://purl.org/spar/deo

described using *AMO*, the *Argument Model Ontology*[18], that implements Toulmin's model of argumentation [23][19] in OWL.

The *textual semantics*, i.e. the very message contained in a piece of text, is the final step in the definition of the semantics of a piece of text. For instance, the formal description of a claim needs to be expressed in such a way as to represent as faithfully as possible the meaning of the claim itself. Since each document expresses content in domains that are specific of the topic of the paper, we cannot provide an encompassing ontology to express claims. In some cases, the claim of an argument can be encoded through using a simple model, e.g. DBPedia.

Finally, a document takes also part to a *citation network* with its cited documents, in particular taking into account the *reasons* for particular citations – e.g. to express qualification of or disagreement with the ideas presented in the cited paper – which may effect the evaluation of a citation network itself. Using CiTO, the *Citation Typing Ontology*[20] [16], we provide descriptions of the nature of the citations.

4 Application of the theory

In this section we provide an answer to the question we introduced in Section 1 – how can the theory of semantic lenses be used to extend effectively semantic publishing capabilities of publishers?

We look at this issue from two orthogonal points of view: (i) identifying the actors involved in the process and (ii) presenting a tool to help readers to focus on distinct aspects of the same document so as to benefit from 'lenses-based' semantic annotations.

4.1 Authoring Semantic Lenses

The application of any particular lens to a document is an authorial operation in the sense that is an act involving individuals acting as agents, responsible for the choice of determined semantic interpretations on a document or its content. Although it seems to be necessary to have authors involved in the application of semantic lenses, thus tracking the provenance of semantic assertions of an enriched document, it may be more difficult and even unclear to understand the possible relationship between the authorship of semantic lenses and the actors involved in that authorship. Semantic Publishing involves different actors of

[18] AMO: http://www.essepuntato.it/2011/02/argumentmodel

[19] Toulmin proposed that arguments are composed of statements having specific argumentative roles: the *claim* (a fact that must be asserted), the *evidence* (a foundation for the claim), the *warrant* (a statement bridging from the evidence to the claim), the *backing* (credentials that certifies the warrant), the *qualifier* (words or phrases expressing the degree of certainty of the claim) and the *rebuttal* (restrictions that may be applied to the claim).

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

the publication chain [22] – such as authors, reviewers, editors and publishers – who may be responsible for the application of particular kinds of metadata rather than others. Within the semantic lenses domain, it is quite important to identify how all these actors are involved in the application of semantic lenses. Of course, there is no clear-cut answer to this question, but based on our own experience in field-testing the application of lenses, we find reasonable to suggest some guidelines, beginning by considering how much the original authors of the document might be involved in the generation of semantic lenses. In Table 1 we summarise our own findings and recommendations about the involvement of the authors or other possible actors that might intervene on each semantic lens.

Table 1. Summary of suggested involvement in the authoring of lenses.

Semantic lens	**Author involvement**	**Other actors usually involved**
Semantics	Highly recommended	Publisher: can specify additional semantics to text (e.g., the abstract). Proof reader: detects errors prior to final publication and can propose appropriate changes. Reader: can provide a semantic interpretation of author's text such as in form of nanopublication.
Argumentation	Highly recommended	Reviewer: can suggest different way of presenting and defending a claim.
Citation	Recommended	Publisher: can expand the citation network of the document according to implicit and/or inferable relations, e.g., links to related papers and so on. Reader: can link the article in unpredictable way according to his/her own interests, also with auxiliary application such as CiteULike Reviewer: can propose additional citation links between the document in consideration and related materials according to particular reasons
Rhetoric	Recommended	Editor: can provide semantics of the particular rhetoric organisation of sections so as to conform the document in consideration to the proposed organisation of a particular journal. Reader: can enhance particular blocks of text so as to make explicit the perceived rhetoric of such a text, e.g. for future searches.
Structure	Limited	Publisher: can suggest and/or apply a different structural organisation of the text according to publishing formats and needs.
Publication context	Not required	Editor and publisher: provide contextual information about the actual venue where the document was published and how it appears within bibliographic reference lists of other papers.
Contributions and roles	Recommended	Publisher: can complete the information provided by authors about their contributions and roles within the document in consideration.
Research context	Highly recommended	Funding agency and institution: can provide additional metadata to describe their involvement related to the document and, thus, to increase their visibility within the Web of Data.

Even if we have broadly identified author's involvement and other actors in semantic lenses applications, the time when one can apply these lenses can vary. On the one hand, the timeframe for the application of the context-specific lenses relates to several aspects that may be gathered only after the document publication (e.g., the publication venue, the DOI, etc.). On the other hand,

according to the other content-specific lenses, there is the possibility to apply them within the same timeframe of the document creation, since the author's involvement would be more straightforward. As a result the information would arguably be far more accurate than a post-hoc application.

However, the above ideal approach does not address some fundamental technical issues. First, it supposes that authors already know how to apply semantic lenses, or could become quickly familiar with semantic lenses, their definitions, and concepts and meanings encoded by the ontologies used. In addition, the application of semantic lenses requires a good amount of technical knowledge, which is an unreasonable expectation for non-experts. In the next section we propose a solution for helping users understanding semantic lenses.

4.2 Through A Lens

The knowledge of the languages used to represent lens-related semantic data is crucial to understand and use semantic lenses appropriately. This knowledge seems to be the most significative obstacle to a wide adoption of semantic lenses, since several actors (e.g. publishers, readers, authors) may not be experts of such semantic technologies. A common solution is to hide the intrinsic complexities of such technologies behind an interface that allows anyone (even the non-expert) to use a tool like semantic lenses in an easy way. To this end, we developed a prototypical HTML interface to scholarly papers called *TAL* (*Through A Lens*), which enables the navigation of a text document on which semantic lenses have been applied to make explicit all the corresponding information. As input, TAL takes an EARMARK representation [8] of a document – we use an HTML version of [14] in the online-available prototype[21] – properly enriched with lens-related semantic assertions as shown in Section 3. The production of annotated documents is not simple. EARMARK includes a Java API on top of which we are developing sophisticated editors. At this stage, we used that API to annotate the sample document. Further developments on the authoring of semantic-lenses-enabled documents are still needed. TAL generates an HTML page with the article and some tools enabling a quick and smart navigation.

Argumentation index. This index is generated from semantic data related to the *argumentation* lens. It lists all the argumentations of the document, making possible to click on each claim within this index to scroll the document down to where the sentence of the claim is written and to show up the related argumentative components (*evidences*, *warrants*, *backings*, *qualifiers* and *rebuttals*). Figure 2.A shows a TAL screenshot with this summary. Claim seven is expanded, others are left unexpanded. Each type of component of a claim (e.g. Evidence, Warrant, Backing, etc.) is explicitly labeled, and coloured in a way to be immediately distinguishable from other types.

Rhetoric Denotation. Labels are placed at the beginning of each paragraph to mean its rhetoric function according to data related to the *rhetoric* lens. Figure 2.B contains the rhetoric denotation of a paragraph.

[21] http://www.essepuntato.it/lisc2014/LensedMika.html

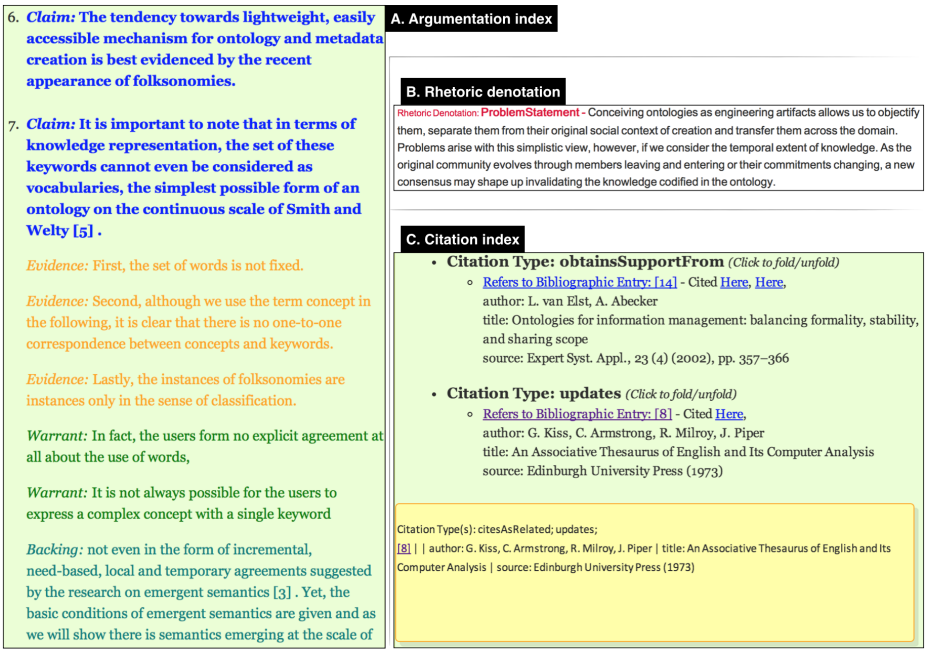


Fig. 2. Three TAL screenshots showing: (A) the argumentation index, (B) a rhetoric denotation of a paragraph, and (C) the citation index with the tooltip box.

Citation index. This index is the counterpart of the argumentation index, but realised over the citation lens. The purpose is to give an interactive table of content for the whole set of citations made by the document, and to offer a level of readability and interactivity similar to the one seen in the argumentation index, by explicitly showing all the citations within the text, grouped by their related CiTO properties and ordered by frequency in the document, together with pointers to their occurrences within the text. An example is shown in Figure 2.C. The position and the way to open the citation index is the same of the argumentation one. Once it expands, the summary reveals a first list of CiTO properties. This list is ordered by frequency of use within the document. Clicking on a property, a nested sub-list is unfolded with the references to all citation items exhibiting that property. To each item is associated a summary of the bibliographic reference information originally contained within the text, together with pointers to both the complete bibliographic reference, as well as anchor links to each occurrence of the citation within the document.

Tooltip box. A yellow box, shown in Figure 2.C, is placed on the right side of the document content. It will be used to show additional information about in-line references (such as the factual or rhetoric reason of citations) and claims (such as the rhetoric denotation of paragraphs containing them) when hovering them with the mouse pointer. All the information visualised in the box

are generated starting from semantic data related to the lenses *argumentation*, *citation* and *rhetoric*.

5 Experiment and evaluation

At this stage of the development of the TAL prototype, we undertook user testing on it, not solely to gather data about its usability and effectiveness, but mostly to probe if the road we had undertaken in order to make available our set of lens browsing features might be potentially promising. We asked 9 subjects with different backgrounds (Ph. D. students and people working in publishing houses) to perform three unsupervised tasks (max. 5 minutes per task), involving navigation of Mika's paper [14] through TAL. There were no "administrators" observing the subjects while they were undertaking these tasks. All the subjects were volunteers who responded to personal e-mails. When prototype development will be over, we plan to execute further user tests, including comparative ones, and with a larger user base.

The tasks given to the subjects are shown in Table 2. This set of tasks was designed to exploring the TAL capabilities in enabling an intuitive and useful navigation of papers. The test session was structured as follows. Firstly, as a warm-up task, we asked subjects to use TAL to find the paragraph containing the second claim and to write down all the citations in that paragraph, explaining also the reason for the citation (max. 5 minutes). Then, as the real test, we asked subjects to complete the three tasks listed in Table 2 using TAL (max. 5 minutes per task). Finally, we asked subjects to fill in two short questionnaires, one multiple choice and the other textual, to report their experience of using TAL to complete these tasks (max. 10 minutes). All the questionaries and all the outcomes of the experiments are available online[22].

Table 2. The three tasks subjects performed in the user testing session.

Task 1	Write down all the reasons why the document cites the reference [8]
Task 2	Write down the evidences of the claim "It is important to note that in terms of knowledge representation, the set of these keywords cannot even be considered as vocabularies, the simplest possible form of an ontology on the continuous scale of Smith and Welty [5]"
Task 3	Write down the (first words of the) paragraphs containing statements of the problems discussed in the paper

Out of 27 tasks in total (3 tasks given to each of 9 subjects), 20 were completed successfully (i.e., the right answers were given), while 7 had incorrect or incomplete answers, giving an overall success rate of 74%. The 20 successes were distributed as follows: 5 in Task1, 9 in Task2 and 6 in Task3.

The usability score for TAL was computed using the *System Usability Scale* (*SUS*) [2], a well-known questionnaire used for the perception of the usability of a

[22] http://www.essepuntato.it/lisc2014/questionaries

system. In addition to the main SUS scale, we also were interested in examining the sub-scales of pure *Usability* and pure *Learnability* of the system, as proposed recently by Lewis and Sauro [13]. As shown in Table 3, the mean SUS score for TAL was 70 (in a 0 to 100 range), surpassing the target score of 68 to demonstrate a good level of usability [20]. The mean values for the SUS sub-scales Usability and Learnability were 69.44 and 72.22 respectively.

Table 3. SUS values and related sub-measures.

Measure	Mean	Max. value	Min. value	S. deviation
SUS value	70	95	50	13.58
Usability	69.44	93.5	53.13	12.18
Learnability	72.22	100	37.5	24.83

6 Conclusions

Modern publishers are now approaching digital publishing from a semantic perspective, making steps towards *semantic publishing*. In this paper we introduce a way to use *semantic lenses* [17] to semantically enhance a published journal article. In addition, we also introduced *TAL*, a prototypical application we developed as proof of concept of the use of semantic lenses in a real-case scenario, that enables the navigation and understanding of a scholarly document through these semantic lenses. Although TAL is still a prototype rather than a complete application, the outcomes reported from the user testing session were positive and very encouraging. In the future we plan to extend TAL so as to handle additional ways of navigation according to all the eight lenses introduced, as well as to produce semantic assertions according to each lens through automatic or semi-automatic approaches, as already proposed for the structural lens [6, 7] and the citation lens [4].

References

1. Belhajjame, K., Zhao, J., Garijo, D., Hettne, K. M., Palma, R., Corcho, O., ... Goble, C. A. (2014). The Research Object Suite of Ontologies: Sharing and Exchanging Research Data and Methods on the Open Web. The Computing Research Repository. http://arxiv.org/abs/1401.4307
2. Brooke, J. (1996). SUS: a "quick and dirty" usability scale. Usability Evaluation in Industry: 189–194. ISBN: 978-0748404600
3. D'Arcus, B., Giasson, F. (2009). Bibliographic Ontology Specification. Specification Document, 4 November 2009. http://bibliontology.com/specification
4. Di Iorio, A., Nuzzolese, A. G., Peroni, S. (2013). Towards the automatic identification of the nature of citations. Proceedings of SePublica 2013. http://ceur-ws.org/Vol-994/paper-06.pdf

5. Di Iorio, A., Nuzzolese, A. G., Peroni, S., Shotton, D., Vitali, F. (2014). Describing bibliographic references in RDF. Proceedings of SePublica 2014. http://ceur-ws.org/Vol-1155/paper-05.pdf
6. Di Iorio, A., Peroni, S., Poggi, F., Vitali, F. (2014). Dealing with structural patterns of XML documents. Journal of the American Society for Information Science and Technology, 65 (9): 1884–1900. DOI: 10.1002/asi.23088
7. Di Iorio, A., Peroni, S., Poggi, F., Vitali, F., Shotton, D. (2013). Recognising document components in XML-based academic articles. In Proceedings of DocEng 2013: 181–184. DOI: 10.1145/2494266.2494319
8. Di Iorio, A., Peroni, S., Vitali, F. (2011). A Semantic Web Approach To Everyday Overlapping Markup. Journal of the American Society for Information Science and Technology, 62 (9): 1696–1716. DOI: 10.1002/asi.21591
9. Groza, T., Moller, K., Handschuh, S., Trif, D., Decker, S. (2007). SALT: Weaving the claim web. Proc. of ISWC 2007: 197–210. DOI:10.1007/978-3-540-76298-0_15
10. IFLA Study Group on the Functional Requirements for Bibliographic Records (1998). Functional Requirements for Bibliographic Records (FRBR), Final Report. http://archive.ifla.org/VII/s13/frbr/frbr_current_toc.htm
11. Lauritsen, M., Gordon, T. F. (2009). Toward a general theory of document modeling. Proceedings of ICAIL 2009: 202–211. DOI:10.1145/1568234.1568257
12. Lebo, T., Sahoo, S., McGuinness, D. (2013). PROV-O: The PROV Ontology. W3C Recommendation, 30 April 2013. http://www.w3.org/TR/prov-o/
13. Lewis, J. R., Sauro, J. (2009). The Factor Structure of the System Usability Scale. Proceedings of HCD 2009: 94–103. DOI: 10.1007/978-3-642-02806-9_12
14. Mika, P. (2005). Ontologies are us: A unified model of social networks and semantics. Web Semantics, 5 (1): 5–15. DOI: 10.1016/j.websem.2006.11.002
15. Peroni, S., Gangemi, A., Vitali, F. (2011). Dealing with Markup Semantics. Proceedings of i-Semantics 2011: 111–118. DOI: 10.1145/2063518.2063533
16. Peroni, S., Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. Web Semantics 17: 33–43. DOI: 10.1016/j.websem.2012.08.001
17. Peroni, S., Shotton, D., Vitali, F. (2012). Faceted documents: describing document characteristics using semantic lenses. Proceedings of DocEng 2012: 191–194. DOI: 10.1145/2361354.2361396
18. Peroni, S., Shotton, D., Vitali, F. (2012). Scholarly publishing and the Linked Data: describing roles, statuses, temporal and contextual extents. Proceedings of i-Semantics 2012: 9–16. DOI: 10.1145/2362499.2362502
19. Pettifer, S., McDermott, P., Marsh, J., Thorne, D., Villeger, A., Attwood, T. K. (2011). Ceci n'est pas un hamburger: modelling and representing the scholarly article. Learned Publishing, 24 (3): 207–220. DOI: 10.1087/20110309
20. Sauro, J. (2011). A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices. ISBN: 9781461062707
21. Schneider, J., Groza, T., Passant, A. (2013). A review of argumentation for the Social Semantic Web. In Semantic Web 4 (2): 159–218. DOI: 10.3233/SW-2012-0073
22. Shotton, D., Portwin, K., Klyne, G., Miles, A. (2009). Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article. PLoS Computational Biology, 5 (4): e1000361. DOI: 10.1371/journal.pcbi.1000361
23. Toulmin, S. (1959). The uses of argument. ISBN: 0521827485

Clustering Citation Distributions for Semantic Categorization and Citation Prediction

Francesco Osborne[1], Silvio Peroni[2,3], Enrico Motta[1]

[1] Knowledge Media Institute, The Open University, Milton Keynes, UK
{francesco.osborne,e.motta}@open.ac.uk
[2] Department of Computer Science and Engineering, University of Bologna, Bologna, Italy
silvio.peroni@unibo.it
[3] STLab, Institute of Cognitive Sciences and Technologies, CNR, Rome, Italy

Abstract. In this paper we present i) an approach for clustering authors according to their citation distributions and ii) an ontology, the *Bibliometric Data Ontology*, for supporting the formal representation of such clusters. This method allows the formulation of queries which take in consideration the citation behaviour of an author and predicts with a good level of accuracy future citation behaviours. We evaluate our approach with respect to alternative solutions and discuss the predicting abilities of the identified clusters.

Keywords: Semantic Web, Research Data, Bibliometric Data, Expert Search, Hierarchical Clustering, Data Mining, OWL, RDF, SPARQL, BiDO

1 Introduction

Exploring and analysing scholarly data [1] help to understand the research dynamics, forecast trends and derive new knowledge, which can be effectively represented by semantic technologies. Within this context, two important tasks are:

1) classifying authors according to a variety of semantic categories in order to facilitate querying, sharing and reusing such data in different context;
2) forecasting their career trends, allowing us to estimate their future citation behaviour.

In this paper we will present an innovative approach to address both tasks by exploiting author citation distributions.

Most of today systems for the exploration of academic data offer citations or citations-based indexes (e.g., h-index, g-index) as ranking metrics and provide interesting visualizations of citation distributions. However, they do not exploit many interesting features which can be derived by the analysis of citation distributions, such as: 1) the trend of the distribution within a certain time interval (e.g., it is steadily rising), 2) the timing of possible acceleration/deceleration (e.g., it started to rise much faster in the last 3 years), 3) the slope of the citation curve (e.g., every year it gains 20% more citations than the year before), 4) the shape of the citation curve (e.g., it is growing according to a logarithmic function), and 5) the estimated citation behaviour in the following years (e.g., authors with a similar pattern usually receive 200±50 citation in their 8^{th} career years).

These features can support formulating queries that take in consideration the diachronic citation behaviour of authors. Examples are: “find all PhD students working in Semantic Web who exhibit a possible rising star pattern”, “find all the senior researchers who in their young years exhibited the same citation pattern as author X” or “find all the postdoc working in UK whose citations exhibit a positive trend in the last two years and are rising exponentially”.

Analysing the citation distributions can also foster a better understanding of the dynamics of an author career, since it makes possible to categorize different kinds of patterns and to study how they evolve. Moreover, it can allow us to forecast the future citation behaviour of research communities or organizations by studying the patterns of their members.

In this paper we present an approach for clustering authors according to their citation distributions, with the aim of extracting useful semantic information and producing statistical evidence about the potential citation behaviour of specific categories of researchers. In addition, we introduce an ontology, i.e., the *Bibliometric Data Ontology* (*BiDO*), which allows an accurate representation of such clusters (and their intended semantics) according to specific categories.

The rest of the paper is organized as follows. In Section 2, we discuss existing approaches for clustering authors and predicting future citations. Section 3 describes our approach for clustering authors’ citation distributions, while Section 4 illustrates BiDO and introduces the steps for associating the identified clusters to ontological categories. In Section 5, we evaluate our approach versus alternative solutions and discuss the predictive abilities of the identified clusters. Finally, in Section 6, we summarize the key contributions of this paper and outline future directions of research.

2 Related Work

Classifying entities associated to a time series is a common task that is traditionally addressed with a variety of clustering techniques [2]. Citation distributions and their mathematical properties have been carefully analysed in a number of empirical studies (e.g., [3]). However, while academic authors are often classified by community detection and clustering algorithms with the aim of identifying different kinds of research communities [4,5], no current model exploits clusters of citation distributions to classify researchers according to the features described earlier and estimate their future citation behaviour.

In the past, several works have been published about the identification of the factors that allow the prediction of future citations. Their analyses, and the related statistical models and machine learning techniques proposed for such predictions, are usually performed according to specific hypotheses: taking into consideration only articles of high-rated journals of a certain discipline; analysing only particular kinds of articles (e.g., clinical articles); choosing only multidisciplinary journals so as to increase the coverage (and the variability) of the research communities involved; and

so forth[2]. As a result, different starting hypothesis gave rise to different (even contrasting) discriminating factors and prediction models.

However, most of these works agree on the existence of two different and complementary kinds of factors:

- *intrinsic* factors, i.e., those related with the qualitative evaluation of the content of articles (quality of the arguments, identification of citation functions, etc.);
- *extrinsic* factors, i.e., those referring to quantitative characteristics of articles such as their metadata (number of authors, number of references, etc.) and other contextual characteristics (the impact of publishing venue, the number of citation received during time, etc.).

The use of intrinsic factors data can be very effective but also time consuming. They can be gathered manually by humans, e.g., through questionnaires to assess the intellectual perceptions of an article (as in peer review processes). For instance, in [7] the authors show how the editor's and reviewer's ratings (in the context of the *Journal of Cardiovascular Research*, http://cardiovascres.oxfordjournals.org) are good predictors of future citations.

The data of some intrinsic factors, such as the identification of citation functions (i.e., author's reasons for citing a certain paper), can also be gathered automatically with the aim of being used to provide alternative metrics for assessing or predicting the importance of articles through machine learning techniques (cf. [8]), probabilistic models (cf. [9]), and other architectures based on deep machine reading (cf. [10]),

However, these approaches use extrinsic factors, rather than intrinsic ones, for the analysis of the importance of articles, because of the time-consuming nature of the latter ones and the quick availability (usually at publication time) of most of the extrinsic-based data. In [11], Didegah and Thelwall investigate the extrinsic factors that better correlate with citation counts, identifying three factors as the best ones for such prediction: the impact factor of the journals where articles have been published, the number of references in articles, and the impact of the papers that have been cited by the articles in consideration. Other extrinsic factors identified in other studies are article length (in terms of printed pages) [12], number of co-authors [13], rank of author's affiliation [13], number of bibliographic databases in which a journal was indexed [14], proportion of the journal articles published that had been judged of high quality by some authoritative source [14], and price index [6]. Slightly different kinds of extrinsic factors were considered in Thelwall *et al.*'s work on altmetrics [15]. The authors analysed eleven different altmetrics sources and found that six of them were good predictors of future citations (i.e., tweets, Facebook posts, Nature research highlights, blog mentions, mainstream media mentions and forum posts).

3 Clustering Citation Distributions

In this section, we will present our approach for detecting clusters of researchers who share a similar citation distribution. We want to identify clusters characterized by citation distributions which represent the typical patterns of some categories of

[2] A good literature review of a large number of such approaches is available in [6].

authors, so that each cluster will suggest a common future behaviour. More formally, we want to subdivide the authors in sets, in such a way that the population of each set will remain homogenous with respect to the number of citations collected in the following years, i.e., the members of each cluster will have a similar number of citations also in the future.

Our approach takes as input the citation distributions of authors in a certain time interval and returns 1) a set of clusters with centroids that describe the most typical citation patterns, 2) a matrix associating each author with a number of clusters via a membership function, and 3) a number of statistics associated to each cluster for estimating the evolution of the authors in that cluster.

We cluster the citation distributions by exploiting a bottom-up hierarchical clustering algorithm. The algorithm takes as input a matrix containing the distance between each couple of entities and initially considers every entity as a cluster. It then computes the distance between each of the clusters, joining the two most similar clusters at each iteration. We adopt a single-linkage strategy by estimating the distance between two clusters C_1 and C_2 as the shortest distance between a member of C_1 and a member of C_2. The algorithm stops when it reaches a certain distance threshold *t*.

To obtain cluster sets that are fit for our purpose we must thus define accordingly 1) the metric to compute the distance between each couple of citation distributions and 2) a method to decide the threshold *t*.

It is possible to measure the distance between two time series by means of metrics such as the Euclidean distance or cosine similarity. Unfortunately both of these solutions have some shortcomings in this case. In fact, when using the Euclidean distance, covariates with the highest variance will drive the clustering process: a threshold value that allows clustering distributions of a certain scale (e.g., 200 citations) will also merge together perfectly valid clusters of minor scale (e.g., 20 citations). The distance based on the cosine similarity (e.g., the inverse minus one) will solve this problem since it is scale-invariant; unfortunately it would also cluster together distributions of completely different scale but with the same shape (e.g., [1,1,2] and [100,100,200]). Let us assume a couple of citation distributions *A* and *B* having both a total of *n* citations, and a different couple of them *C* and *D* with *m* citation each, *C* having the same distribution as *A*, and *D* the same as *B*. We want a distance that will yield *dis{A,B} = dis{C,D}* (avoiding the covariate with the highest variance to drive the clustering) and also *dis{A,C} > 0* (making scale a feature), and furthermore can be calculated incrementally (thus sparing processing time by stopping the computation over a threshold). A simple way to satisfy these three requirements makes use of a Euclidean distance normalized with the number of total citations of both distributions (similarly to [16]):

$$EU_n = \left(\frac{\sum_{i=0}^{n} (x_i - y_i)^2}{\sum_{i=0}^{n} (x_i)} + \frac{\sum_{i=0}^{n} (x_i - y_i)^2}{\sum_{i=0}^{n} (y_i)} \right) / 2 \qquad (1)$$

where x_i and y_i are the number of citations of the two distributions in the *i*-th year.

We also want to choose a threshold value *t* that will maximize the homogeneity of the cluster populations in the following years. We compute the homogeneity of a population with respect to citations using the Median Absolute Deviation (MAD). MAD is a robust measure of statistical dispersion [17] and it is used to compute the

variability of an univariate sample of quantitative data. It was first used by Gauss for determining the accuracy of numerical observations and it is defined as the median of the absolute deviations from the original data's median:

$$MAD = median_i\left(\left|x_i - median_j\left(x_j\right)\right|\right) \quad (2)$$

The procedure for computing the MAD consists in calculating the median of the n original data ($x_1,x_2,\ldots,x_j,\ldots,x_n$), computing the differences between each one of the n original values x_i and the median of the whole data distribution and finally computing the median of the previous differences. We preferred MAD to different solutions, such as standard deviation, for its robustness. In fact, standard deviation is too much influenced by outliers such as a few authors with a very high number of citations. Hence, we estimate the quality of a set of clusters in a certain year by computing the weighted average of their MAD:

$$MAD_{av} = \frac{\sum_{i=0}^{n}(MAD(c_i)\cdot dim(c_i))}{dim(c_i)} \quad (3)$$

where $MAD(c_i)$ and $dim(c_i)$ are respectively the MAD and the number of authors associated with the i-th cluster.

We set the threshold t by running the hierarchical algorithm with different t values and then selecting the threshold which yields clusters with the lowest average MAD_{av} in the following n years (n=10 in the herein presented evaluation). For characterizing completely the author space we compute the clusters for different intervals of time, e.g., 1-5, 1-10 and 1-15 career years, using a significant author sample (e.g., 5000). We then compute the memberships of all authors in our dataset with the centroids of the resulting clusters, so as to determine exactly how much a specific author is similar to each cluster centroid. For associating authors to clusters, we adopt the well known membership formula of the Fuzzy C-Mean algorithm [18], that is:

$$mem_k(x) = \frac{1}{\sum_{i=0}^{n}\left(\frac{dis(center_k,x)}{dis(center_j,x)}\right)^{2/(m-1)}} \quad (4)$$

where $mem_k(x)$ is the membership value of author x with cluster k, $dis(center_i, x)$ the distance between x and the centroid of cluster i, and m is a constant for modulating the level of cluster fuzziness (m=2 in the prototype).

Finally we analyse the distribution of each cluster population with respect to the number of citations received in the following years, in order to extract statistical evidence about their future behaviour. As mentioned before, standard deviation is severely influenced even by few outliers, making it hard to use the mean on the full population as a predictor. Hence, for each year we automatically select a percentage p of the population (e.g., 90%) in the most populated area of the distribution and compute its interval of citations (e.g., 40-80), mean (e.g., 45) and standard deviation (e.g., 14). Technically, we do so by computing the number of authors who fall into different ranges of citations, ordering those categories in decreasing order and then selecting the authors from subsequently smaller categories until the percentage of authors selected is equal to p. The citation interval, mean and standard deviation of this sample produce accurate, intuitive and statistically sound predictions which are more resilient to outliers.

Intuitively, some categories of authors are too mundane to suggest a common future behaviour, and may be used only for classification purposes. Hence, in this phase we care especially about the "uncommon signature" that points to particularly homogenous population of authors. Figure 1 shows the distributions of authors in their seventh career year associated to some clusters detected by analysing their first five career years (the dashed line refers to the overall distribution). Clusters *C29* and *C30* are associated with a very specific citation patterns and thus their distributions have a small kurtosis and point to two narrow categories of authors who normally receive a relatively low number of citations. Clusters *C25* and *C28* are also quite homogeneous and represent two distinct populations of more frequently cited authors. Naturally, the homogeneity of the population associated with a cluster will decrease in the following years and so will the accuracy of the predictions.

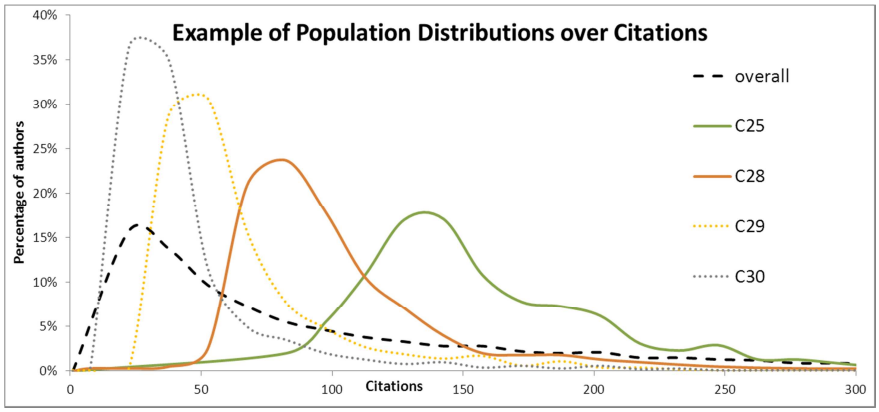


Figure 1. Percentage of authors vs. number of their citations in their 7th career year. The clusters were derived by the citations received over the first 5 career years.

4 An ontology for describing bibliometric data

Having a model developed according to a well-known format (such as OWL) for enabling the classification of authors and journals according to bibliometric data is crucial to allow one to query, share and reuse such data in different context, e.g., for providing smart visualisation of bibliometric data for sense-making activities and for enabling automatic reasoning on them.

However, bibliometric data are not simple objects, since they are subject to the simultaneous application of different variables. In particular, one should take into account at least:

- the *temporal association* of such data to entities, in order to say that a particular value, e.g., the fact that an article has been cited 42 times, was associated to such article only for a time period;
- the particular *agent who provided* such data (e.g., Google Scholar, Scopus, our algorithm), in order to keep track of the way data evolve in time according to particular sources;

- the *characterisation* of such data in at least two different kinds, i.e., numeric bibliometric data (e.g., the standard bibliometric measures such as h-index, journal impact factor, citation count) and categorial bibliometric data (so as to enable the description of entities, e.g., authors, according to specific descriptive categories).

The *time-indexed value in time* (*TVC*) ontology design pattern [19] seems to be a good starting model for the development of an ontology for bibliometric data, since TVC's entities enable the precise description of all the aforementioned variables: time, responsible agent and kinds of data.

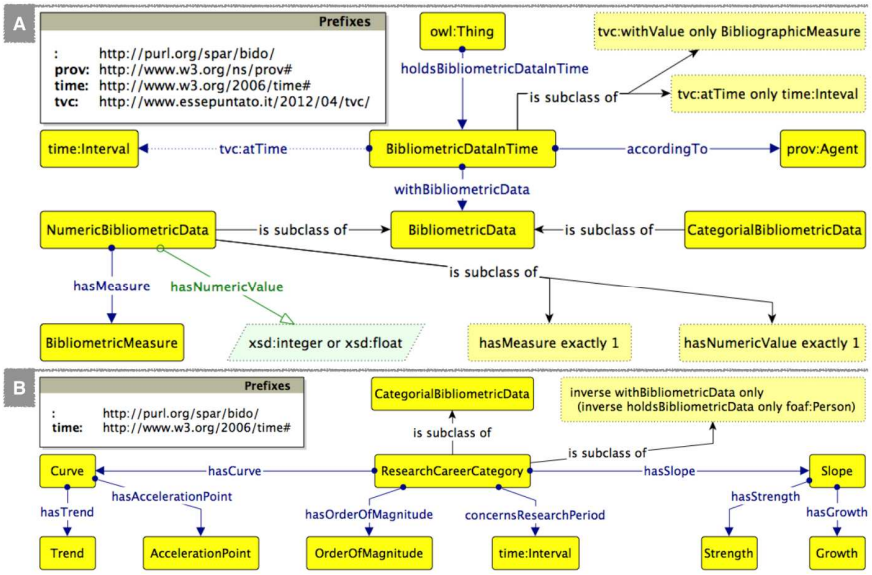


Figure 2. *A:* the core module of BiDO, describing generic bibliometric data with their characterising variables. *B:* the module modelling a particular kind of categorial bibliometric data, i.e., the research career categories, according to the main dimensions used by the algorithm in Section 3.

Starting from TVC, we have created the *Bibliometric Data Ontology* (*BiDO*, available at http://purl.org/spar/bido), i.e., a modular OWL 2 ontology that allows the description of bibliometric data of people, articles, journals, and other entities described by the SPAR Ontologies (http://purl.org/spar) in RDF.

The core module of the ontology, shown in Fig. 2.A, allows us to describe any entity and the related bibliometric data (through the property *holdsBibliometricDataInTime*) at a certain time (i.e., *tvc:atTime*, a property defined by the imported TVC ontology for specifying temporal instants or intervals) and according to a certain agent (through the property *accordingTo*, which is a sub-property of *prov:wasAttributedTo* and allows us to indicate the agent responsible for such bibliometric data). In addition, BiDO imports PROV-O [20] for adding provenance data about the activities related to the creation of such bibliometric data.

Two alternative kinds of bibliometric data are specifiable (through the property *withBibliometricData*) in BiDO: numeric and categorial bibliometric data. Numeric bibliometric data are those characterised by a certain integer or float value related to a particular bibliometric measure. Some of these measures – i.e., *h-index*, *author citation count*, *e-index*, and *journal impact factor* – are available in a particular module of BiDO responsible for describing the most common bibliometric measures.

We have developed an additional module of BiDO that extends the class *CategorialBibliometricData* of the core module with specific categories describing the research career of people, in order to address the mapping of the clusters identified by the algorithm presented in Section 3 with specific facets. As shown in Fig. 2.B, these facets are described by the class *ResearchCareerCategory*, which is characterised by four specific dimensions that have been used by our algorithm to cluster citation data:

- the *research period* considered, i.e., the interval of research years that the algorithm is taking into consideration (e.g., the first 5/10 years);
- the *curve*, i.e., the specific shape proper to the clusters identified by the algorithm, which is characterised by a trend (flat/increasing/decreasing) and, in the latter two cases, by an acceleration or deceleration point (none or premature, median, overdue acceleration/deceleration);
- the *slope* of such curve, in terms of strength (low/moderate/high) and kind of growth (linear/polynomial/exponential/logarithmic);
- the *order of magnitude*, which categorises the number of citations received in the considered period according to a uniform model of common-sense estimation [21], which describes intervals of half-order of magnitude – i.e., "[0,1)", "[1,3)", "[3,9)", "[9,27)", "[27,81)", "[81,243)", "[243,729)", etc.

The combinations of all these values related to the aforementioned dimensions have been used to define all the possible descriptive categories of research career of people as instances of the class *ResearchCareerCategory*.

Even if we did not define a particular category for each cluster found by the algorithm – rather, more clusters can be described by the same category –, we have defined an algorithmic procedure to determine the association between the cluster centroids and the categories described by the ontological model. For instance, let us consider the centroid "[31.3, 46.1, 52.8, 55.3, 60.8]"[3] of one of the clusters detected by our algorithm according to the first 5 years of research career. The related dimensions are identified in the following way:

- *order of magnitude*: we sum the values of the cluster centroid and select the interval containing such sum, i.e., "[243,729)";
- *curve trend*: the linear regression of the centroid is calculated, and then its slope is divided by the mean of all the centroid values. If the result of such division is greater than 0.05, then we have an increasing trend (which is the case of our example, since that value is 0.14), if it is less than -0.05 we have a decreasing trend, otherwise we have what we can approximately consider a flat trend;

[3] The five values of the centroid identify the number of citations that have been received during the five years of the research period considered.

- *curve acceleration*: the ratio of the slopes of the linear regressions of series *k-n* and *1-k* (for each *k* between 2 and *n* - 1, where *n* is length of the list of values defining a cluster centroid) is calculated, in order to identify in which year (i.e., *k*) the acceleration or deceleration (this is the case of our example) happens, if any. Then, the acceleration/deceleration is considered premature if $k \leq \lceil n/3 \rceil$ (as in our example), overdue if $k \geq \lceil 2n/3 \rceil$, and median otherwise;
- *slope strength*: the linear regression of the centroid is calculated, its slope is divided by the mean of all the centroid values, and then we calculate the absolute value *s* of this division. We say that the slope strength is low if *s<0.25* (as in our example), high if *s>0.45*, and moderate otherwise;
- *slope growth*: by means of the least squares method, we create the four functions (one linear, one polynomial, one exponential and one logarithmic) that best match with the cluster centroid. Then we compare the centroid data with such functions through Wilcoxon's non-parametric test for matched data and choose the best fitting function (logarithmic in our example).

Following these steps, the example cluster we considered is mapped in the following category:

```
:increasing-with-premature-deceleration-and-low-logarithmic-slope-in-[243,729)-5-
years-beginning a :ResearchCareerCategory ;
  :hasCurve [ a :Curve ;
    :hasTrend :increasing ; :hasAccelerationPoint :premature-deceleration ] ;
  :hasSlope [ a :Slope ; :hasStrength :low ; :hasGrowth :logarithmic ] ;
  :hasOrderOfMagnitude :[243,729) ;
  :concernsResearchPeriod :5-years-beginning .
```

Thus, combining the results of our clustering algorithm with BiDO it is possible to associate authors with specific categories describing their research career as follows:

```
ex:john-doe :holdsBibliometricDataInTime [
  a :BibliometricDataInTime ;
  tvc:atTime [ a time:Interval ; time:hasBeginning :2014-07-11 ] ;
  :accordingTo [ a fabio:Algorithm ;
    frbr:realization [ a fabio:ComputerProgram ] ] ;
  :withBibliometricData
    :increasing-with-premature-deceleration-and-low-logarithmic-slope-in-
[243,729)-5-years-beginning .
```

The RDF descriptions of such bibliometric data make easier to query them with standard languages such as SPARQL, in order to retrieve, for instance, all the authors that in the first 5 years of their research career had a citation behaviour pattern like that described by the aforementioned category.

5 Evaluation

We evaluated our method on a dataset of 20000 researchers working in the field of computer science in the 1990-2010 interval. This dataset was derived from the database of Rexplore [1], a system that combines statistical analysis, semantic technologies and visual analytics to provide support for exploring scholarly data, and integrates several data sources (Microsoft Academic Search, DBLP++ and DBpedia).

In particular we wanted to show that the normalized Euclidean distance introduced in Section 3 works better than other choices for the task of clustering citation distributions. Hence, we compared three metrics: the normalized Euclidean distance (label NEU), the Euclidean distance (EU) and the distance based on the cosine similarity (CO). We measured the quality of the produced set of clusters in a certain year by their MAD_{av}, as in Formula (3).

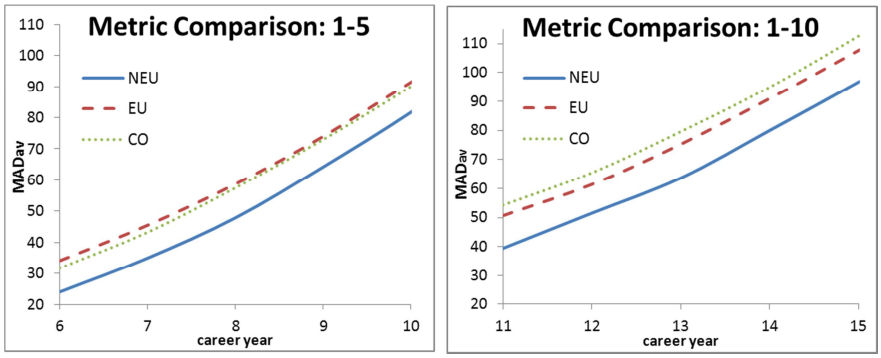


Figure 3. Comparison between NEU, EU and CO applied on the first five and ten career years according to their MAD_{av} in the following five years.

Figure 3 shows the performance of the three techniques when clustering the first five and ten career years. In all cases the normalized version of the Euclidean distance performs much better than the other solutions, being characterized by a smaller MAD_{av} value, e.g., a smaller degree of dispersion. CO performs slightly better than EU in the 1-5 years interval while EU performs better than CO in the 1-10 years interval. Analogous results were obtained by considering the weighted average of standard deviation rather than MAD_{av}.

Career year	C18 (1.4%)		C22 (2.5%)		C25 (2.7%)		C28 (2.3%)		C29 (8.8%)	
	range	mean±s.d.	range	mean±s.d.	range	mean±s.d.	range	mean±s.d.	range	mean±s.d.
6	420-800	567±98	160-280	209±34	100-180	129±25	60-100	72±14	40-60	39±9
7	440-960	610±120	160-320	225±45	100-200	138±30	60-120	79±18	40-80	45±14
8	440-1020	650±137	160-400	246±58	100-260	158±45	60-160	90±26	40-100	50±18
9	440-1260	699±186	160-440	269±74	100-340	187±68	60-200	104±37	40-120	57±25
10	480-2940	751±411	160-500	292±85	100-400	211±82	60-280	125±57	40-160	68±35
11	480-2480	826±336	180-660	331±112	100-520	241±100	60-540	155±103	40-200	82±47
12	480-3520	914±467	180-860	370±151	100-640	270±126	60-440	166±96	40-260	97±60

Table 1. Range of citations and mean citations in subsequent career years predicted with 75% accuracy for authors associated with clusters detected in the 1-5 career year interval. In parenthesis the percentage of authors in each cluster.

Our approach yields a number of clusters with different prediction capabilities. We can suggest a narrower or larger interval of predicted citations for increasing or lowering the precision of our predictions. Table 1 shows some example of predictions that yield 75% accuracy. For example we are able to suggest with 75% precision that 2.5% authors in Computer Science associated with cluster *C22* will have 225±45 average citations in their seventh career years (with a minimum number of citations equal to 160 and a maximum one equal to 320).

The left panel of Figure 4 shows the citation distributions of the centroids of the cluster in Table 1 and the algorithm predictions. Even if the predictions become less accurate in time, however they still can give a fair idea of the kind of potential citation behaviour of the authors. Moreover, these predictions are particular valuable for forecasting the future citation behaviour of an organization or research communities. In fact, while it is relatively hard to foresee a single author's citation behaviour (e.g., she/he may be an outlier), it is much easier to compute the predicted citations of a group of authors since in a large sample statistical fluctuations have a smaller weight.

Finally, the right panel of Figure 4 shows the evolution of some the main clusters in terms of average citations of the associated authors. We can notice that our approach allows a very good coverage of the possible career trajectories, from the most modest to the outstanding ones. This variety of patterns allow also for a very fine-grained semantic classification of researcher careers.

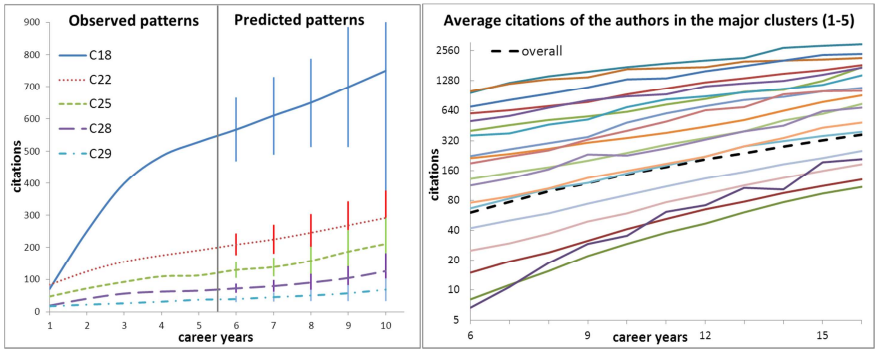


Figure 4. Left Panel: the citation distributions of the centroids of the clusters in Table 1 and the resulting predictions (the error bars represent the standard deviations of the predicted citations). Right Panel: the evolution in term of average number of citations of the authors associated to the main clusters in the 1-5 interval.

6 Conclusion

In this paper, we presented a novel approach for clustering author's citation distributions, with the aim of 1) classifying authors with a variety of semantic facets, and 2) forecasting the citation behaviour of categories of researchers. We also introduced the Bibliometric Data Ontology, a.k.a. BiDO, which is an OWL ontology that allows an accurate representation of such semantic facets describing people's research careers. In addition, we showed that our approach outperforms other solutions in terms of population homogeneity and is able to categorize a variety of career trajectories, some of which allow predicting future citations with fair accuracy.

For the future we plan to augment the clustering process with a variety of other features (e.g., research areas, co-authors), to extend BiDO in order to provide a semantically-aware description of such new features, and to make available a triplestore of bibliometric data linked to other datasets such as Semantic Web Dog Food and DBLP.

References

1. Osborne, F., Motta, E., Mulholland, P.: Exploring Scholarly Data with Rexplore. In Proceedings of the ISWC 2013: 460-477. (2013)
2. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On clustering validation techniques. Journal of Intelligent Information Systems, 17(2-3), 107–145. (2001)
3. Redner, S.: How popular is your paper? An empirical study of the citation distribution. The Physics of Condensed Matter Journal, 4(2), 131–134. (1998)
4. Ding, Y.: Community detection: topological vs. topical. Journal of Infometrics, 5(4). (2011)
5. Osborne, F., Scavo, G., Motta, E.: Identifying diachronic topic-based research communities by clustering shared research trajectories. In The Semantic Web: Trends and Challenges (pp. 114-129). Springer International Publishing. (2014)
6. Onodera, N., & Yoshikane, F.: Factors affecting citation rates of research articles: Factors Affecting Citation Rates of Research Articles. Journal of the Association for Information Science and Technology. (2014)
7. Opthof, T.: The significance of the peer review process against the background of bias: priority ratings of reviewers and editors and the prediction of citation, the role of geographical bias. Cardiovascular Research, 56(3), 339–346. (2002)
8. Teufel, S., Siddharthan, A., Tidhar, D.: Automatic classification of citation function. In Proceedings of the EMNLP 2006: 103–110. Stroudsburg, Pennsylvania, USA. (2006)
9. Dietz, L., Bickel, S., Scheffer, T.: Unsupervised prediction of citation influences. In Proceedings of ICML 2007: 233–240. (2007).
10. Di Iorio, A., Nuzzolese, A. G., Peroni, S.: Towards the automatic identification of the nature of citations. In Proceedings of SePublica 2013. (2013)
11. Didegah, F., Thelwall, M.: Determinants of research citation impact in nanoscience and nanotechnology. Journal of the American Society for Information Science and Technology, 64(5): 1055–1064. (2013)
12. Falagas, M. E., Zarkali, A., Karageorgopoulos, D. E., Bardakas, V., Mavros, M. N.: The Impact of Article Length on the Number of Future Citations: A Bibliometric Analysis of General Medicine Journals. PLoS ONE, 8(2), e49476. (2013)
13. Antonakis, J., Bastardoz, N., Liu, Y., Schriesheim, C. A.: What makes articles highly cited? The Leadership Quarterly, 25(1), 152–179. (2014)
14. Lokker, C., McKibbon, K. A., McKinlay, R. J., Wilczynski, N. L., Haynes, R. B.: Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study. BMJ, 336(7645), 655–657. (2008)
15. Thelwall, M., Haustein, S., Larivière, V., Sugimoto, C. R.: Do Altmetrics Work? Twitter and Ten Other Social Web Services. PLoS ONE, 8(5), e64841. (2013)
16. Ning, T.: Computing correlation integral with the Euclidean distance normalized by the embedding dimension. In Proceedings of ICSP 2008: 2708–2712. (2008)
17. Hoaglin, D. C., Mosteller, F., Tukey, J. W.: Understanding robust and exploratory data analysis (Vol. 3). New York: Wiley. (1983)
18. Bezdek, J. C., Ehrlich, R., Full, W.: FCM: The fuzzy c-means clustering algorithm. Computers and Geosciences, 10(2), 191–203. (1984).
19. Peroni, S., Shotton, D., Vitali, F.: Scholarly publishing and linked data: describing roles, statuses, temporal and contextual extents. In Proceedings of i-Semantics 2012: 9–16. (2012).
20. Lebo, T., Sahoo, S., McGuinness, D.: PROV-O: The PROV Ontology. W3C Recommendation, 30 April 2013. World Wide Web Consortium. (2013)
21. Hobbs, J. R., Kreinovich, V.: Optimal choice of granularity in commonsense estimation: Why half-orders of magnitude? International Journal of Intelligent Systems, 21(8), 843–855. (2006)

SMART Protocols: SeMAntic RepresenTation for Experimental Protocols

Olga Giraldo[1], Alexander García[2], and Oscar Corcho[1]

[1] Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
{ogiraldo, ocorcho}@fi.upm.es
[2] Linkingdata I/O LLC, Fort Collins, Colorado, USA
alexgarciac@gmail.com

Abstract. Two important characteristics of science are the "*reproducibility*" and "*clarity*". By rigorous practices, scientists explore aspects of the world that they can reproduce under carefully controlled experimental conditions. The clarity, complementing reproducibility, provides unambiguous descriptions of results in a mechanical or mathematical form. Both pillars depend on well-structured and accurate descriptions of scientific practices, which are normally recorded in experimental protocols, scientific workflows, etc. Here we present SMART Protocols (SP), our ontology-based approach for representing experimental protocols and our contribution to clarity and reproducibility. SP delivers an unambiguous description of processes by means of which data is produced; by doing so, we argue, it facilitates reproducibility. Moreover, SP is thought to be part of e-science infrastructures. SP results from the analysis of 175 protocols; from this dataset, we extracted common elements. From our analysis, we identified document, workflow and domain-specific aspects in the representation of experimental protocols. The ontology is available at http://purl.org/net/SMARTprotocol

Keywords: experimental protocol, ontology, in vitro workflow, reproducibility.

1 Introduction

Scientific experiments often bring together several technologies at in vivo, in vitro and sometimes in silico levels. Moreover, the biomedical domain relies on complex processes, comprising hundreds of individual steps usually described in experimental protocols. An experimental protocol is a sequence of tasks and operations executed to perform experimental research. The protocols often include equipment, reagents, critical steps, troubleshooting, tips and all the information that facilitates reusability. Researchers write the protocols to standardize methods, to share these documents with colleagues and to facilitate the reproducibility of results.

Although reproducibility, central to research, depends on well-structured and accurately described protocols, scientific publications often lack sufficient information when describing the protocols that were used. For instance, there is

4th Workshop on Linked Science 2014- Making Sense Out of Data (LISC2014),
19th or 20th October 2014, Riva del Garda, Trentino, Italy

ambiguity in the terminology as well as poor descriptions embedded within a heterogeneous narrative. There is the need for a unified criterion with respect to the syntactic structure and the semantics for representing experimental protocols.

Here we present SMART Protocols (henceforth SP), our ontology-based approach for representing experimental protocols. SP aims to formalize the description of experimental protocols, which we understand as domain-specific workflows embedded within documents. SP delivers a structured workflow, document and domain knowledge representation written in OWL DL. For the representation of document aspects we are extending the Information Artifact Ontology (IAO).[1] The representation of executable aspects of a protocol is captured with concepts from P-Plan Ontology (P-Plan) [1]; we are also reusing EXPO [2], EXACT [3] and OBI [4]. For domain knowledge, we rely on existing biomedical ontologies. SP results from the analysis of 175 experimental protocols gathered from several sources. From this dataset, we extracted common elements and evaluated whether those protocols could be implemented.

Our main assumption is that *"experimental protocols are fundamental information structures that should support the description of the processes by means of which results are generated in experimental research"*. Hence our approach should allow answering questions such as: *Who is the author of the protocol? What is the application of the protocol? What are the reagents, equipment and/or supplies used? What is the estimated time to execute a protocol? Which samples have been tested in a protocol?* This paper is organized as follows: Section 2 presents related works, Section 3 describes the methodology stages to develop the SP ontology, section 4 shows the results and ontology evaluation. Finally Section 5 provides discussion and conclusions.

Related Work

In an effort to address the problem of inadequate methodological reporting, the MIBBI[2] project brings under one umbrella most of these projects. The ISA-TAB also illustrates work in this area; it delivers metadata standards to facilitate data collection, management and reuse [5].

The Ontology for Biomedical Investigations (OBI)[3] aims to model the design of investigations, including the protocols, materials used and the data generated. OBI has key classes for the description of experiments, namely: `obi:investigator`, `obi:instrument`, `obi:biomaterial entity`. The generic ontology of scientific experiments (EXPO)[4] aims to formalize domain-independent knowledge about the planning, execution and analysis of scientific experiments. This ontology includes the class `expo:ExperimentalProtocol` and defines some of its properties: `expo:has_applicability`, `expo:has_goal`,

[1] https://code.google.com/p/information-artifact-ontology/
[2] http://mibbi.sourceforge.net/portal.shtml
[3] http://obi-ontology.org/page/Main_Page
[4] http://expo.sourceforge.net/

`expo:has_plan`. EXACT suggests a meta-language for the description of experiment actions and their properties.

Recently, PLOS ONE in collaboration with Science Exchange and Figshare launched "The Reproducibility Initiative".[5] This project aims to help scientists to validate their research findings. The Research Object initiative[6] aims to deliver a model to represent experimental resources; this model facilitates accessibility, reusability, reproducibility and also a better understanding of in silico experiments.

Publishers are also actively addressing the problem of experimental reproducibility; F1000Research,[7] an open science journal, suggests data preparation guidelines to capture the processes and procedures required to publish scientific dataset. The Force 11 initiative,[8] a community of researchers addressing issues in scholarly communication, has published a set of metadata standards for biomedical research. These standards focus on three recommendations: Gene accession numbers, organism identification and reagent identification. Vasilevsky et al., [6] recently published a study addressing the issue of material resource identification in biomedical literature. Interestingly, the results indicated that 54% of the resources are not uniquely identifiable in publications.

Unlike other approaches, the SP ontology provides a formalized representation of the domain that is not sufficiently covered by other ontologies. For instance, SP-document delivers a structured vocabulary representing a specific type of document, a protocol. This vocabulary includes rhetorical components (e.g. introduction, materials, and methods); it also has information like application of the protocol, advantages and limitations, list of reagents, critical steps. In addition, The formalization of instructions in the protocol, or steps, is covered in SP-workflow by the class `p-plan:Step.` The order in which these steps should be executed is captured by the property `bfo:isPrecededBy`. Inputs and outputs from each step are represented by the class `p-plan:Variable`.

2 Methodology

For designing SP, we followed practices recommended by the NeOn methodology [7]. Also, we carefully considered the experience reported by García [8]; for example, we used conceptual maps to better understand the correspondences, relations and feasible hierarchies in the knowledge we were representing. In addition, concept maps proved to be simpler for exchanging models with domain experts. The stages and activities we implemented throughout our ontology development process are illustrated in Fig 1.

[5] http://blogs.plos.org/everyone/2012/08/14/plos-one-launches-reproducibility-initiative/
[6] http://www.researchobject.org/
[7] http://f1000research.com/data-preparation
[8] https://www.force11.org/Resource_identification_initiative

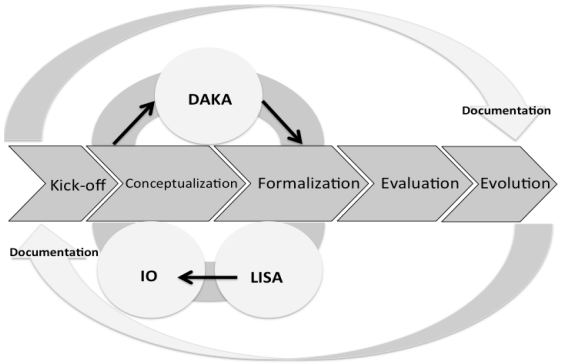


Figure 1. Methodology used to develop SMART Protocols.

2.1 Kick-off

In this stage we gathered motivating scenarios, competency questions, and requirements. We focused on the functional aspects we wanted the ontology to support. Competency questions were specified with domain experts, some of them are presented below: *i) Who is the author of the protocol?, ii) What is the application of the protocol?, iii) What is the provenance of the protocol?, iv) Who are the manufacturer and catalog number of reagents, equipment or supplies used?, v) What is the estimated time to execute a protocol?, vi) Which samples have been tested with a protocol?, vii) What are the critical steps, tips or troubleshooting of a protocol? viii) What are the basic steps of protocols in molecular biology?*

2.2 Conceptualization and formalization

In this stage we identified reusable terminology from other ontologies; supporting activities throughout this stage we used BioPortal.[9] We also looked into minimal information standards,[10] guidelines and vocabularies representing research activities [9-11]. Issues about axioms required to represent this domain were discussed and tested in Protégé v. 4.3; during the iterative ontology building, classes and properties were constantly changing. We identified three main activities throughout this stage, namely:

1. **Domain Analysis and Knowledge Acquisition, DAKA:** from the journals we worked with, protocols and guidelines for authors were analyzed; theory vs. practice was our main concern, *What information elements were required? Was there any relation between terminology from ontologies and these set of requirements from journals?* We also manually verified if published protocols were following the guidelines, if not, *What was missing?* Throughout this activity

[9] http://bioportal.bioontology.org/
[10] https://www.force11.org/node/4145

we were also analyzing existing ontologies and minimal information standards against published protocols. DAKA was facilitated because the knowledge engineer, namely Olga Giraldo, was also a domain expert with over ten years working in a laboratory of biotechnology. 35 domain experts were active participants in the development of SP; they were responding surveys, attending workshops, assisting in the definition of competency questions and scenarios of use. They were also validating the terminology and the relations.

We manually reviewed 175 published and non-published protocols from domains like biotechnology, virology, biochemistry and pathology. The non-published protocols (75 in total) were collected from four laboratories located at International Center for Tropical Agriculture (CIAT).[11] The published protocols (open access protocols in plant biology) were gathered from 9 repositories: Biotechniques,[12] Cold Spring Harbor Protocols (CSH Protocols),[13] Current Protocols (CP),[14] Genetics and Molecular Research (GMR),[15] Journal of Visualized Experiments (JoVE),[16] Protocol Exchange (PE),[17] Plant Methods (PM),[18] Plos One (PO)[19] and Springer Protocols (SP)[20] (Table 1).

Table 1. Repositories and number of protocols analyzed.[21]

Repository	CP	JoVE	PE	PM	CSH	Bio Tech.	GMR	PO	SP
No. of protocols	25	21	13	12	9	6	5	5	4
Total	**100**								

2. **Linguistic and Semantic Analysis, LISA:** this is the most complex activity throughout our development process. We identified linguistic structures that authors were using to represent actions; we needed to understand how instructions were organized. We were interested in understanding how verbs were representing actions, what additional information was there for indicating attributes for actions. By analyzing texts we were also identifying terminology and determining whether these terms were already available in existing ontologies. Minimal information standards were also considered; how could these be used when describing an experimental protocol?

[11] http://ciat.cgiar.org/
[12] http://www.biotechniques.com/protocols/
[13] http://cshprotocols.cshlp.org/
[14] http://www.currentprotocols.com/WileyCDA/
[15] http://www.geneticsmr.com/
[16] http://www.jove.com/
[17] http://www.nature.com/protocolexchange/
[18] http://www.plantmethods.com/
[19] http://www.plosone.org/
[20] http://www.springerprotocols.com/
[21] http://goo.gl/MC4mR9

From our dataset we extracted common elements and evaluated whether those protocols could be implemented. Initially, we focused our analysis on identifying necessary and sufficient information for reporting protocols. From our inspection, we determined workflow aspects in experimental protocols. The sequence of instructions had an implicit order, following the input output structure. Actions in the workflow of instructions were usually indicated by verbs; accurate information for implementing the action implicit in the verb was not always available. For instance, structures such as *"Mix thoroughly at room temperature*", "*Briefly spin the racked tubes*" are common in our dataset. Due to the ambiguity and lack of detailed information for specifying actions in the instructions, it was difficult to understand how could these be implemented. Domain expertise was usually required in order to interpret some of the actions in our dataset.

In addition, we also isolated elements pertaining to domain knowledge as well as document related characteristics. We classified our protocols within 4 groups according to the purpose, namely: i) plant genetic transformation, ii) DNA/RNA extraction and purification, iii) PCR and their variants, iv) electrophoresis and sequencing. Within each group we identified basic steps (or common patterns), which we consider as necessary in the structure of the protocol. For example, we found that a cell disruption step is essential in DNA extraction protocols. We also identified that a digestion reaction (removing the lipid membrane, proteins and RNA) follows and that the DNA precipitation or purification comes at the end of this process.

Variables	Constants
Cell disruption (CD)	First= first step
Digestion reaction (DR)	Second= second step
DNA precipitation or purification (DNAP)	Third= third step

```
dna_extraction_protocol(CD, DR, DNAP):–
     CD= first, DR= second, DNAP= third
```

3. **Iterative ontology building and validation, IO:** as we were gathering information and learning about this domain, we started by building concept maps; these were rapidly mapped to parts of speech from the texts we were analyzing and also to existing ontologies. As concept maps were growing in complexity, number of concepts and relations, we then started to build draft ontologies –baseline ontologies representing specifics from parts of speech we identified. The knowledge engineer conducted the evaluation of the draft ontologies against competency questions. Models were also exchanged with domain experts; the process was iterative and, the models were constantly growing.

By building ontology models as well as by carefully analyzing the information we were gathering from LISA and DAKA activities, we were able to identify the modularity needed to represent experimental protocols. The module SP-document was designed to provide a structured vocabulary of concepts to represent information

for recording and reporting an experimental protocol. The module SP-workflow aims to provide a structured vocabulary of concepts to represent the execution of experimental protocols in life sciences.

2.3 Evaluation

The goal of the evaluation is to determine what the ontology defines, and how accurate these definitions are. Here we follow the activities proposed by Gómez-Pérez et al. [12] for terminology evaluation, which provide the following criteria:

1. **Consistency**. It is assumed that a given definition is consistent if, and only if, no contradictory knowledge may be inferred from other definitions and axioms in the ontology.
2. **Completeness**. It is assumed that ontologies are in principle incomplete [12, 13], however it should be possible to evaluate the completeness within the context in which the ontology will be used. An ontology is complete if and only if: o All that is supposed to be in the ontology is explicitly stated, or can be inferred.
3. **Conciseness.** An ontology is concise if it does not store unnecessary knowledge, and the redundancy in the set of definitions has been properly removed.

According to the criteria for evaluation proposed by Gomez-Perez [12], our ontologies were developed using the OWL-DL because of expressiveness and computational completeness.[22] The Protégé plugin OWLViz[23] was used to visualize and to correct syntactic inconsistencies. The OntOlogy Pitfall Scanner (OOPS),[24] was useful to detect and correct anomalies or pitfalls in our ontologies [14]. In relation to the evaluation of the terminology, we represented the 175 protocols using the SMART Protocol formalism, emphasizing on informative elements. For most of the cases, we had insufficient information from the protocols; domain expertise was therefore required in order to determine what was missing and how to best represent it. We also used surveys[25] in order to determine how complete our model was. As a result from our analysis we proposed a checklist[26] to report experimental protocols in plant biology; 35 domain experts validated this checklist.

2.4 Evolution

At the end of the cycle, new classes, properties and individuals are identified. These are then analyzed against the set to competency questions, existing ontologies, parts of speech and linguistic structures. The model evolves as new knowledge goes through the whole cycle. Although our ontology is a young ontology, we could observe how it evolved in its conceptualization as well as in the explicit specification.

[22] http://www.w3.org/TR/2004/REC-owl-features-20040210/#s1.3
[23] http://protegewiki.stanford.edu/wiki/OWLViz
[24] http://oeg-lia3.dia.fi.upm.es/oops/index-content.jsp
[25] goo.gl/jBHPo
[26] goo.gl/gAVnn

3 The SMART Protocols Ontology

The SMART Protocols approach follows the OBO Foundry principles [15]. Our modules reuse the Basic Formal Ontology (BFO).[27] Also, we reused the ontology of relations (RO) [16] to characterize concepts. In addition, each term from SP is represented by annotation properties imported from OBI Minimal metadata.[28]

An overview of the two modules comprising SP is illustrated in Figure 2. The classes, properties and individuals are represented by their respective labels to facilitate the readability. The prefix indicates the provenance of each term. The ontology describing the experimental protocol as a document is depicted at the top. The class `iao:information content entity` and its subclasses `iao:document`, `iao:document part`, `iao:textual entity` and `iao:data set` were imported from The Information Artifact Ontology (IAO) to represent the document aspects in the protocol. The ontology describing the experimental protocol as a workflow is depicted at the bottom. The representation of executable aspects of a protocol is modeled with the classes `p-plan:Plan`, `p-plan:Step` and `p-plan:Variable` from the P-Plan Ontology (P-Plan).

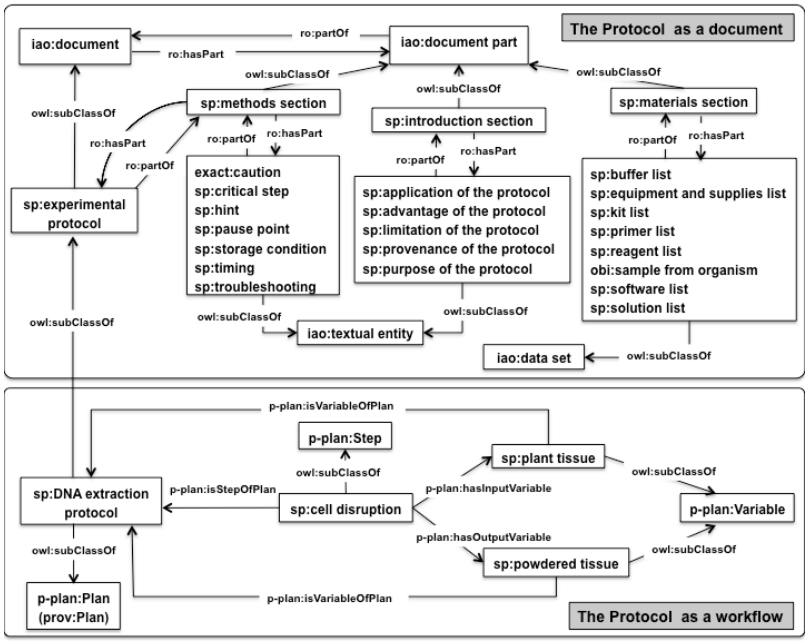


Figure 2. SMART Protocols as an extension of the ontologies IAO and P-Plan. The document aspects in a protocol are captured with IAO. The workflow aspects in a protocol are captured with P-Plan. The terms proposed in SMART Protocols use the sp prefix.

[27] http://www.ifomis.org/bfo/
[28] http://obi-ontology.org/page/OBI_Minimal_metadata

3.1 The protocol as a document

The document module of SMART Protocols reuses classes from CHEBI [17], EXACT, MGED [18] , SO [19], OBI and SNPO.[29] Also, SMART Protocols-document (henceforth SP-document) extends the class `iao:information content entity` proposed by the Information Artifact Ontology (IAO) to represent the experimental protocol as an `iao:document` that has parts, `ro:has_part`, such as `iao:document part` (`iao:author list`, `sp:introduction section`, `sp:materials section` and `sp:methods section`). See the top of Figure 2 for details.

Use Case. SP-document represents information such as, the protocol type, `sp:DNA extraction protocol`; it has a tittle, identified by the property `sp:has title`, it is instantiated by genomic DNA isolation. Also, the author entry, `iao:author identification`, is instantiated by CIMMYT [20]. This protocol is derived, `sp:provenance of the protocol`, from the protocol published by [21] (sp:PNAS 81:8014-8019) and its purpose is instantiated by plant DNA extraction of high quality (Fig. 3).

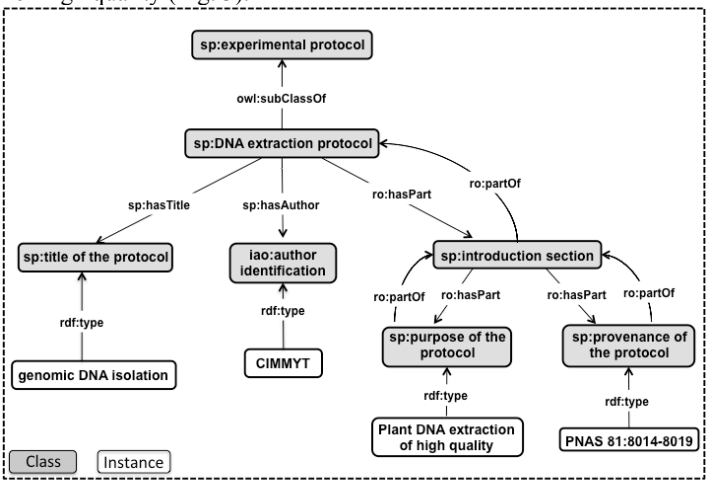


Figure 3. The document aspects in a protocol are captured with IAO. The terms proposed in SMART Protocols use the sp prefix.

3.2 The protocol as a workflow

The workflow module extends the P-Plan Ontology (P-Plan). This ontology was developed to describe scientific processes as plans and link them to their previous executions. In the workflow module of SMART Protocols (henceforth SP-workflow), the experimental protocol, `p-plan:Plan`, is a description of a sequence of

[29] http://www.loria.fr/~coulet/snpontology1.4_description.php

operations, `p-plan:Step`, that includes an input and an output `p-plan:Variable`. In this sense, a protocol is a type of workflow. See the bottom of Figure 2 for details. SP-workflow also reuses classes from CHEBI, MGED, SO, OBI and NPO [22].

Use Case. DNA extraction is a procedure frequently used to collect DNA for subsequent molecular or forensic analysis. DNA extraction includes 3 basic `p-plan:Steps`: i) cell disruption or cell lysis, ii) Digestion reaction (in this step, contaminants such as lipid membrane, proteins and RNA are removed from the DNA solution), and iii) DNA purification. Each one of these steps may include different protocols (or `p-plan:Plans`) to be executed. For example, the step `sp:cell disruption` or cell lysis may be achieved by chemical and physical methods - blending, grinding or sonicating the sample. Also, the ontology considers that each step is executed following a predetermined order. For instance, according to the protocol published by CIMMYT, the cell disruption by lyophilization and grinding has an input variable, `p-plan:hasInputVar`, as well as `sp:plant tissue`; it also has an output, `p-plan:hasOutputVar`, and `sp:powdered tissue`. The next step, `sp:digestion reaction`, has as input the output of the immediately previous step, `sp:powdered tissue`, and as output `sp:digested contaminant`. The last one, `sp:DNA purification` has as input `sp:digested contaminant`, and as output `obi:DNA extract` (Fig. 4).

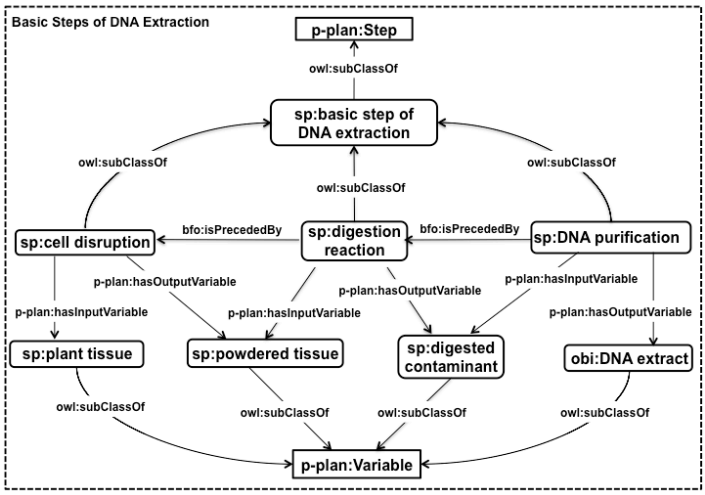


Figure 4. Extending the P-plan ontology to represent experimental protocols in life sciences. The sp prefix indicates the terms proposed by the SMART Protocols ontology.

4 Discussion and Conclusions

Science has, among other, two important characteristics, reproducibility and clarity.[30] Clarity provides unambiguous descriptions for results in a mechanical or mathematical form. The lack of clarity about "*how to do* or *how to execute*" an experimental procedure hinders the reproducibility and impedes comparing results across related experiments. SMART Protocols addresses clarity by formalizing the objects that should go together with actions. Besides, SP reuses and extends minimal information standards, incorporating these structures within the representation of experimental protocols. By delivering a semantic and syntactic structure SP also facilitates reproducibility.

Our ontology-based representation for experimental protocols is composed of two modules, namely SP-document and SP-workflow. In this way, we represent the workflow, document and domain knowledge implicit in experimental protocols. Our work extends IAO and P-Plan ontology. Actions, as presented by [3] are important descriptors for biomedical protocols; however, in order for actions to be meaningful, attributes such as measurement units and material entities (e.g., sample, instrument, reagents, etc.) are also necessary.

Formalizing workflows has an extensive history in Computer Science; not only in planning but also in execution –as in Process Lifecycle Management and Computer Assisted Design/Computer Assisted Manufacturing. We have considered some of these principles for representing workflow aspects in protocols just as we have reused knowledge formalisms like the P-Plan Ontology. By formalizing the workflow implicit in protocols the execution can be ontologically represented in a sequential manner that is intelligible by humans and processed by machines.

Modularization, as it has been implemented in SP, facilitates managing the ontology. For instance, the workflow module can easily be specialized with more specific formalisms so that robots can process the flow of tasks to be executed. The document module facilitates archiving; the structure also allows to have fully identified reusable components. By combining both modules we are delivering a self-describing document.

References

1. Garijo, D. and Y. Gil. *Augmenting PROV with Plans in P-PLAN: Scientific Processes as Linked Data*. in *2nd international Workshop on Linked Sicence 2012 - Tackling Big Data (LISC2012), in conjunction with 11th International Semantic Web Conference (ISWC2012)*. 2012. Boston, MA: Springer-Verlag.
2. Soldatova, L.N. and K.R. D., *An ontology of scientific experiments.* journal of the royal society interface, 2006. **3**(11): p. 795–803.

[30] http://www.aaas.org/page/so-can-science-explain-everything

3. Soldatova, L.N., et al., *The EXACT description of biomedical protocols.* Bioinformatics, 2008. **24**(13): p. i295-303.
4. Courtot, M., et al. *The OWL of Biomedical Investigations* in *OWLED workshop in the International Semantic Web Conference (ISWC)*. 2008. Karlsruhe, Germany.
5. Sansone, S.A., et al., *Toward interoperable bioscience data.* Nat Genet, 2012. **44**(2): p. 121-6.
6. Vasilevsky, N.A., et al., *On the reproducibility of science: unique identification of research resources in the biomedical literature.* PeerJ, 2013. **1**: p. e148.
7. Suárez-Figueroa, M.C., *Ontology engineering in a networked world*. 2012, Berlin ; New York: Springer. xii, 444 p.
8. Garcia-Castro, A., *Developing Ontologies in the Biological Domain* in *Institute for Molecular Bioscience*2007, University of Queensland: Queensland. p. 275.
9. Zimmermann, P., et al., *MIAME/Plant - adding value to plant microarrray experiments.* Plant Methods, 2006. **2**: p. 1.
10. Bustin, S.A., et al., *The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments.* Clin Chem, 2009. **55**(4): p. 611-22.
11. Gibson, F., et al., *Guidelines for reporting the use of gel electrophoresis in proteomics.* Nat Biotech, 2008. **26**(8): p. 863-864.
12. Gomez-Perez, A., *Evaluation and assessment of knowledge sharing technology*, in *Towards Very Large Knowledge Bases: Knowledge Building & Knowledge Sharing*, N.J.I. Mars, Editor. 1995, IOS Press: Amsterdan, The Netherlands. p. 289-296. .
13. Gómez-Pérez, A., M. Fernández-López, and O. Corcho, *Ontological engineering: with examples from the areas of knowledge management, e-commerce and the Semantic Web*. 2004: Springer.
14. Poveda-Villalón, M., M. Suárez-Figueroa, and A. Gómez-Pérez, *Validating Ontologies with OOPS!*, in *Knowledge Engineering and Knowledge Management*, A. ten Teije, et al., Editors. 2012, Springer Berlin Heidelberg. p. 267-281.
15. Smith, B., et al., *The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.* Nature Biotechnology, 2007. **25(11)**: p. 1251 - 1255.
16. Smith, B., et al., *Relations in biomedical ontologies.* Genome Biology, 2005. **6**(5): p. -.
17. de Matos, P., et al., *Chemical Entities of Biological Interest: an update.* Nucleic Acids Research, 2010. **38**: p. D249-D254.
18. Stoeckert Jr, C.J. and H. Parkinson, *The MGED ontology: a framework for describing functional genomics experiments.* Comparative and Functional Genomics, 2003. **4**: p. 127-132.
19. Mungall, C.J., C. Batchelor, and K. Eilbeck, *Evolution of the Sequence Ontology terms and relationships.* J Biomed Inform, 2011. **44**(1): p. 87-93.
20. CIMMYT, *Laboratory Protocols: CIMMYT Applied Molecular Genetics Laboratory*, 2005, CIMMYT: Mexico, D.F. p. 102.
21. Saghai-Maroof, M.A., et al., *Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics.* Proc Natl Acad Sci U S A, 1984. **81**(24): p. 8014-8.
22. Thomas, D.G., R.V. Pappu, and N.A. Baker, *NanoParticle Ontology for cancer nanotechnology research.* J Biomed Inform, 2011. **44**(1): p. 59-74.

LinkedPPI: Enabling Intuitive, Integrative Protein-Protein Interaction Discovery

Laleh Kazemzadeh[1], Maulik R. Kamdar[1], Oya D. Beyan[1], Stefan Decker[1], and Frank Barry[2]

[1] Insight Center for Data Analytics, National University of Ireland, Galway
{laleh.kazemzadeh,maulik.kamdar,oya.beyan,stefan.decker}@deri.org
[2] Regenerative Medicine Institute, National University of Ireland, Galway
frank.barry@nuigalway.ie

Abstract. Understanding the dynamics of protein-protein interactions (PPIs) is a cardinal step for studying human diseases at the molecular level. Advances in "sequencing" technologies have resulted in a deluge of biological data related to gene and protein expression, yet our knowledge of PPI networks is far from complete. The lack of an integrated vocabulary makes querying this data difficult for domain users, whereas the large volume makes it difficult for intuitive exploration. In this paper we employ Linked Data technologies to develop a framework 'LinkedPPI' to facilitate domain researchers in integrative PPI discovery. We demonstrate the semantic integration of various data sources pertaining to biological interactions, expression and functions using a domain-specific model. We deploy a platform which enables search and aggregative visualization in real-time. We finally showcase three user scenarios to depict how our approach can help identify potential interactions between proteins, domains and genomic segments.

Keywords: Protein-Protein Interaction Network, Linked Data, Domain-specific Model, Visualisation

1 Introduction

1.1 Background

The study of biological networks forms the integral core of biomedical research related to human diseases and drug development. The ultimate goal of such studies is to understand the connections between different genes and proteins, how the cell signals propagate across these networks and regulate their functionality. Hence understanding the Protein-Protein Interaction (PPI) networks underlying each such cellular mechanism is important to specifically target the dysfunctional proteins, leading towards the discovery of potential drugs and treatments for diseases. Studying PPI networks helps understand the interconnectedness between different cellular mechanisms and pathways. Biological pathways are not independent of each other, but their interactions are harmonious, which makes them part of a bigger network. Thus it is important to investigate the dynamics of the cell system as a whole.

Human genome contains more than 20000 protein-coding genes which interact tightly in order to regulate various cellular pathways and mechanisms. The major challenge in developing a thorough understanding of these cellular mechanisms and pathways is to complete the PPI network for each mechanism. However, experimental validation of the binary interactions between total number of proteins is an inconceivable task thus computational models can be used to aid the researchers. These models help identify the sequential, structural and physicochemical properties of known interacting protein pairs and highlight the underlying patterns. Researchers then apply these patterns to narrow down the potential interacting partners for any protein(s) under investigation. Therefore wet-lab validation of the hypothesis formed around the predicted links and protein partners is realistic and achievable.

In computational models, experimentally validated PPIs form the backbone of PPI networks, however data pertaining to gene-expression, domain-domain interactions and genomic locations have proved their valuable contribution in inference and prediction of new links between protein pairs [6,19]. Each of these data sources has been published to address specific, albeit very different research problems. Therefore the data representation, data model and formats may vary from one data source to the other. Challenges stemming from the heterogeneity of the data emphasis the need for a framework which can bridge these biologically different concepts in order to highlight and extract the ubiquitous patterns, inconspicuous in the *bigger picture*.

1.2 Motivation

Due to advances in sequencing technologies, enormous amount of experimental data has been generated and stored as independent databases. Databases such as BioGRID [28], HPRD [9], MINT [17] contain the experimentally validated binary interactions, while UniProt [30], Ensembl [8], Entrez-Gene [21] and Gene Ontology [2] offer sequence information, genome localisation and cellular functionality of individual genes and proteins. On the other hand, knowledge bases like Pfam [27] contain information regarding the functional and structural protein subunits (domains).

The main motivation of this work is to provide researchers with a framework which enables them to retrieve the answers to their research questions from these disparate data sources. A researcher interested in the list of protein domains in a specific protein can look up the UniProt website[3] which is extensively rich in protein information. Genomic locations of protein-coding genes are publicly available from several websites such as CellBase [5]. However questions like, *'List of all the proteins which contain the exact or partial set of protein domains?'* or *'What is the relation of a set of interacting proteins and the genomic location of their underlying genes?'* cannot be answered through these websites.

The challenges in the aggregation and exploration of the aforementioned massive biological data sources have sparked the interests of several domain researchers and led them towards the adoption of a new generation of integrative

[3] `http://www.uniprot.org`

technologies, based on Semantic Web Technologies and Linked Data concepts, thus giving birth to Integrative Bioinformatics [25,7].

2 Methods

The final goal of this research is the identification and extraction of potential PPI networks from various publicly available data sources. The core structure of the PPI network consists of proteins and their experimentally proven interactions. Fig. 1 depicts an overview of the LinkedPPI architecture. Following subsections will describe data selection, RDFization and integration methodologies used.

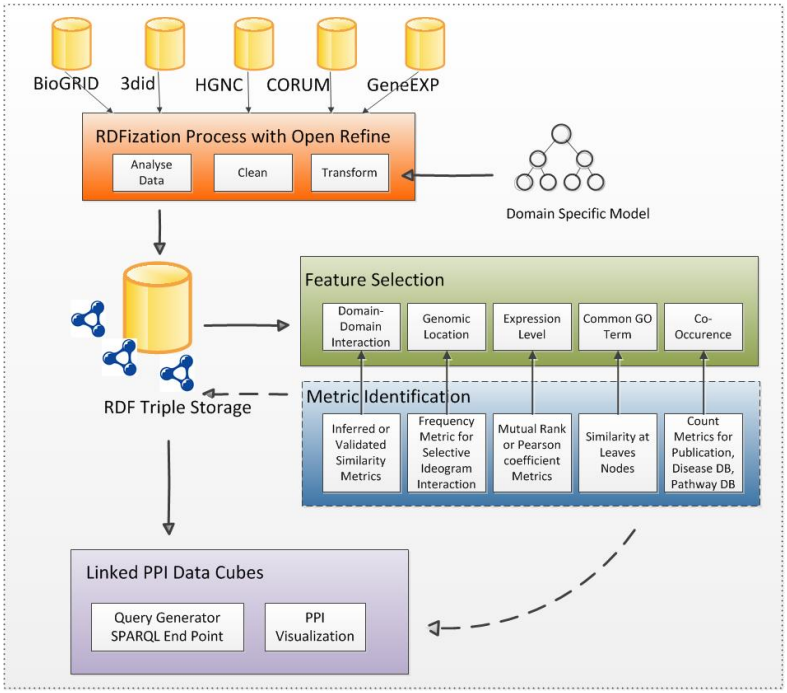


Fig. 1. LinkedPPI Architecture

2.1 Selection of Relevant Data Sources

Validated Interactions: Experimentally validated interactions were retrieved from BioGRID (Biological General Repository for Interaction Datasets), one of the most comprehensive PPI databases [28]. Only physical interactions were included in our work, regardless of their classifications as raw or non-redundant.

Protein Complexes: In most cellular processes proteins act as a complex, instead of binary interactions between a pair of single proteins [1], during the same time and within the same cellular compartment. Such proteins are tightly interacting and play key roles in PPI networks. Elucidation of the dynamics of

PPI networks and functionality of individual proteins can benefit from identification of essential protein complexes, since different subunits contribute to drive a cellular function. In this work, we have used the latest release of CORUM (Comprehensive Resource of Mammalian protein complexes) [24].

Gene Expression: Understanding the correlation between gene-expression networks and protein interaction networks is an ongoing challenge in PPI studies. Proteins coded from co-expressed genes are more likely to interact with each other [10] and there is higher probability that an interacting pair of proteins share cellular functions [3]. We have used the COXPRESdb database [23] which publishes recent gene expression microarray datasets for Human.

Genomic Locations: Neighboring genes show similar expression pattern and are often involved in similar biological functions [22] which suggest that they might share same activation and translation mechanisms. These interactions may not be limited to the adjacent genes but can be long-range interactions to fulfil the cellular functionality [18]. Such evidences encouraged us to introduce a layer for the genomic locations of the protein-coding genes in our framework. We do not define 'genomic location' as the exact start/stop position of genes on a chromosome, but as the Ideogram band in which the genes reside. Ideograms are schematic representations which depict fixed staining patterns on a tightly coiled chromosome in Karyotype experiments. Karyotype describes number of chromosomes, their shape and length and banding patterns of chromosomes in the nucleus. Ideogram data was downloaded from the Mapping and Sequencing Tracks in the Human Genome Assembly (GRCh37/hg19, Feb 2009) at the UCSC Genome Browser[4]. The start/stop coordinates of the genes were retrieved from CellBase [5] and used to determine the genes within each ideogram. HGNC (HUGO Gene Nomenclature Committee) was used to map common genes referenced by different identifiers (Entrez-Gene, Ensembl and UniProt) [26].

Protein Domains: Proteins functionality and their structures are defined by their domain specification. Each protein consists of single or multiple domains, mutual sharing of which may lead to interaction with other proteins. However, identification of domain interactions through experimental validation for all possible protein pairs is an insurmountable task. Therefore domain knowledge bases can shed light on PPIs as well as help identify novel domain-domain interactions. We used 3did (Database of three-dimensional interacting domains) [29] which contains high resolution three-dimensional structurally interacting domains.

Gene Co-occurrence: Studying co-occurrence networks of genes can lead to the prediction of novel PPIs and discovery of hidden biological relations [13]. Previously Kamdar et al. generated co-occurrence scores as a weighted combination of the total number of diseases, pathways or publications in which any two genes occur simultaneously [14].

2.2 LinkedPPI Data Integration

One of the crucial challenges in integrative bioinformatics is the heterogeneous nature of biological data sources. Even though several attempts have been made

[4] `http://genome.ucsc.edu/`

in the standardization of the data through controlled vocabularies and guidelines, various hurdles still need to be surpassed. The proteomic standards initiative-molecular interaction (PSI-MI) is widely accepted by the community for the modelling of biological networks [12]. Even though some of the data sources are represented using the PSI-MI format there is no decipherable interconnectedness between them. To introduce the desired interconnectedness or 'bridges' between these data sources, we decided to use Linked Data technologies.

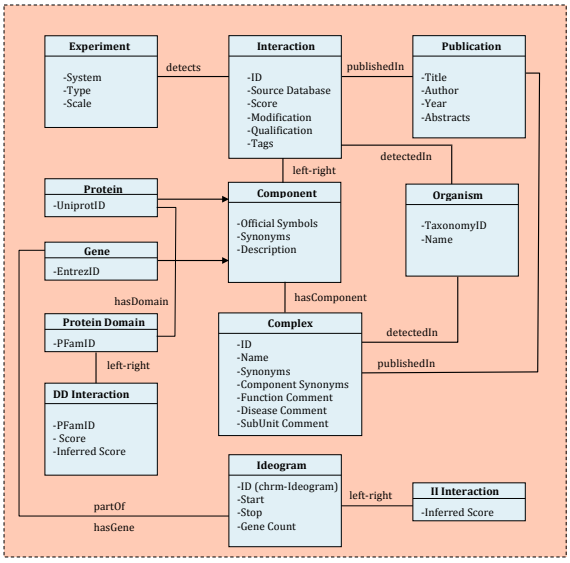


Fig. 2. Class Diagram of LinkedPPI Domain-specific Model

We proposed a simple concise domain model, for the modelling of the PPIs retrieved from BioGRID, complexes from CORUM, protein domains and genomic location. A domain-specific model is beneficial over an extensive well-construed ontology due to the absence of non-domain-specific concepts (`Thing`, `Continuant`, etc.) and is much smaller and self-contained to address a specific problem. Being native to a particular domain (e.g. *Protein-protein interactions*), it serves as an intermediate layer between the user and the underlying data, and enables intuitive knowledge exploration and discovery [15]. Our model comprises 12 concepts, which are termed relevant in this domain, and is shown in Fig. 2. The core concept in this model is a *Component*. A *Component* can either be a *Gene* or a *Protein*. A *Component* can be part of a *Complex* or can interact with another *Component* through an *Interaction*. The *Organism*, in which both the *Interaction* and the *Complex* are detected, is also available as a distinct concept. The *Experiment* concept embodies the attributes related to the experimental system which was used to detect the interaction (e.g. Y2H, AP-MS), the scale of the experiment (high or low-throughput), and whether it is a physical interaction or genetic. *Publication* documents the experiments, and links to resources

described in the PubMed repository. A *Gene* is contained within an *Ideogram*, and *IIinteraction* represents inferred interactions between two ideograms from experimentally validated PPIs. *Domains* associated with protein-coding genes are represented using Pfam IDs and *DDinteraction* models interaction scores between two domains retrieved from 3did and inferred from BioGRID.

Open Refine provides a workbench to clean and transform data and eventually export it in required format. We used its RDF Extension [20] to model and convert the tab-delimited files downloaded from CORUM, BioGRID, 3did and CellBase to RDF graphs and stored them in a local Virtuoso Triple Store[5]. Data from COXPRESdb was already published on the web as RDF, and we re-used their data model and URIs. Similarly, for other data sources we re-used the URIs for the genes, proteins and publications, from those provided by Entrez-gene, UniProt and PubMed. To determine which gene is responsible for the encoding of which protein (mapping between Entrez-Gene ID and UniProt ID), we used the ID mapping table[6] provided by UniProt. One of the major advantages of using this approach was that the mapping also linked the relevant Gene Ontology (GO) [2] terms to the Entrez-Gene ID, thus providing additional information regarding the localisation and function of the specific genes.

3 Results

After RDFization, the BioGRID data source contains around 11 million triples (11357231), which establish 634996 number of distinct interactions between 14135 Human proteins. The data source also links out to 38952 unique PubMed publications documenting these PPIs. The CORUM data source consists of 156364 triples, with 2867 distinct complexes. The 3did data source consists of 320690 triples with 6818 distinct protein domains and 61582 validated and inferred domain-domain interactions. We inferred a total of 13493 interactions between 405 ideograms, referenced through 80092 triples. 60676 mappings were instantiated between the genes and 7750 extracted GO child leaves.

3.1 Search and Visualization

As such, relevant information could be retrieved from the SPARQL Endpoint through the formulation of appropriate queries. However as SPARQL requires a steep learning curve, the non-technical domain user needs intuitive, interactive visualization tools, which aggregate this information from the multiple data sources and summarize it. We devised a PPI Visualization Dashboard[7] based on ReVeaLD (Real-time Visual Explorer and Aggregator of Linked Data) [15] to accommodate our requirements for the search and visual exploration of the LinkedPPI networks. As the user starts typing the official symbol of the desired protein, a list of possible alternatives will be retrieved from the indexed entities. On selection, the entity URI is passed as a parameter through a set of

[5] `http://srvgal78.deri.ie/sparql`

[6] `ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/README`

[7] `http://srvgal78.deri.ie/linkedppi`

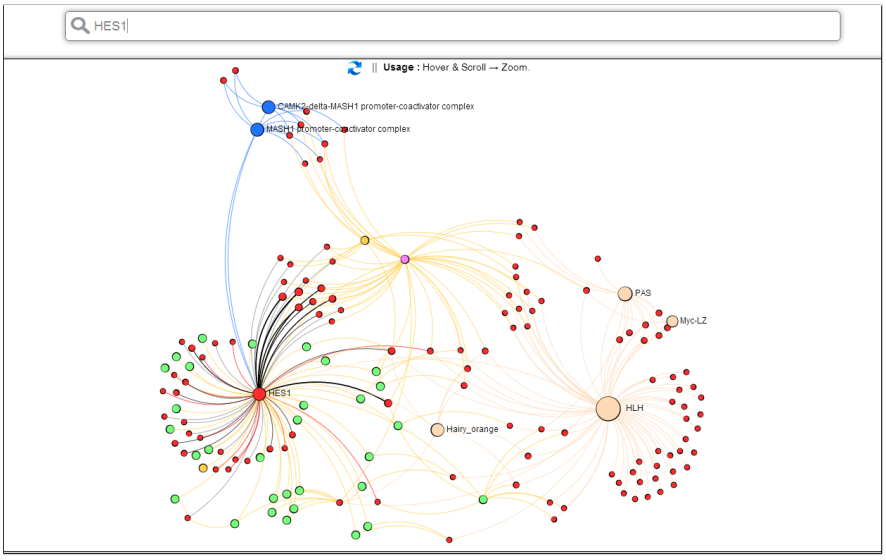


Fig. 3. Searching HES1 protein using the PPI Visualization Dashboard

pre-formulated SPARQL SELECT queries[8] targeting the various data sources. As shown in Fig. 3, the PPI network associated with the searched protein (e.g., HES1 *entrezgene:3280*) is rendered in a force-directed layout. The list of entities retrieved from the data sources are represented as circular nodes, with the size of each node directly proportional to the number of associated nodes. The nodes are rendered using different colors for the sake of visual differentiation - *Red* for *Components* (*Proteins* of BioGRID or *Genes* of COXPRESdb), *Blue* for CORUM *Complexes*, *Light Brown* for 3did *Protein Domains*. The three categories of GO Child Nodes - *Biological Processes*, *Molecular Functions* and *Cellular Components* are displayed using *Green*, *Yellow* and *Purple* colors.

The interactions between different proteins are represented as edges - the color of the edges is directly dependent on whether the associations have been retrieved from BioGRID, COXPRESdb or Co-occurrence Data (*Black*, *Red* and *Purple* respectively). The thickness of *Black* edges depends on the total number of publications which have experimentally validated the underlying interactions. The thickness of the *Red* and *Purple* edges depends on the PCC (Pearson Correlation Coefficient for Gene Expression) and Co-occurrence scores respectively. The *Protein* nodes, which are present in the same complex, possess interacting domains or have underlying coding genes associated to the same GO terms, are not connected directly to each other by edges. They are all connected using similar colored edges to the respective node (complex, domain or GO term), however there may be instances with experimental interactions or co-expression between the connected entities. The resulting network is hence densely clustered, rather

[8] `https://gist.github.com/maulikkamdar/a47fbecddecc6ba4b373`

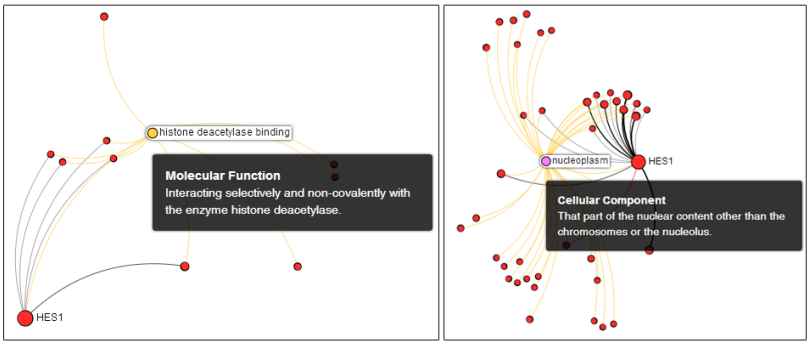


Fig. 4. Subgraph of HES1 PPI network based on GO terms

than a simplistic radial layout of nodes. Hovering over any node highlights subgraph of the network which only displays the first-level connected nodes and their relations (Fig. 4), hence allowing any domain user to intuitively deduce answers to simple questions like, *'Which protein-encoding genes in the network share the same molecular function and have experimental co-expression?'* An information box is also displayed beside the hovered node to show additional information like GO term descriptions, Pfam or PubMed IDs, and PCC scores. Zooming and panning across the visualization is possible using a mouse.

3.2 Use Cases

The following subsections describe three different scenarios depicted in Fig. 5 that our framework could be employed to facilitate extraction of implicit information which can be used as predictors of novel protein-protein interactions. The relevant SPARQL Queries are documented at `http://goo.gl/xesMjR`.

Use Case 1: Extraction of Potential Protein-Protein Interactions Based on the Domain-Domain Interactions. Proteins carry on their functions through their protein domain(s), is a well-known fact. In this scenario we aim to extract possible PPIs based on the known domain-domain interactions. For the sake of simplicity in this use case we assume we are interested in proteins which contain single domains. A researcher has a *protein* in mind for which the sequence specification and domain composition are known. An interesting question might be, list of potential protein partners for this *protein*. Using our framework researcher can retrieve the list of protein pairs in which at least one of the proteins contains the same protein domain as the protein under question. Possible outcomes are: a) the protein under investigation itself shows up in the result set which forms the list of its experimentally validated protein partners. However this could be queried from the BioGRID web site directly. b) List of proteins with one single domain. In this case with a naive and straightforward conclusion, the researcher may accept the list. However in most cases further application of GO enrichment or advanced statistical analysis offer a more concise list but these analyses are beyond the scope of this work. c) The

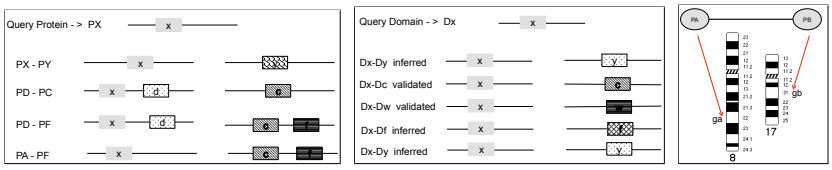
query results to a set of proteins consisting of several domains which requires further statistical or domain expert knowledge refinement. Despite the need of further investigation in such cases, the shortlisted hypothetical interaction partners are expected to be brief to save a tremendous amount of time and effort. The SPARQL Query uses the example of the HES1 (*entrezgene: 3280*) protein. We obtain the list of domains present in HES1 - Hairy_Orange (*pfam:PF07527*) and HLH(*pfam:PF00010*), and the list of proteins (e.g. HEY2) which share these domains, or have domain-domain interactions. We then retrieve validated PPIs in which the protein participates (e.g. HEY2-SIRT1). We can also obtain the PubMed publication documenting each PPI.

Use Case 2: Identification of Potential Domain-Domain Interactions. Protein-protein interactions can be identified experimentally through various types of experiments (e.g: Yeast Two-Hybrid). However it is not possible to identify the interacting domains between two proteins from same experiments and it requires a set of different experiments and protocols. Often protein domains act as signature elements and repeatedly interact with each other within the same organism. Therefore these frequent observations assist in identification of novel domain-domain interactions which is enlightening in identification of latent PPIs. Nevertheless in this work such observations are inferred implicitly from the validated PPI dataset (BioGRID) and require further statistical significance analysis. In our SPARQL Query example, we retrieve the validated and the inferred scores for domain interactions with the HLH domain (*pfam:PF00010*).

Use Case 3: Identification of Selective Interactions between Segments of Human Genome. Human chromosomes are compact in 3D space with each chromosome folding into its own territory [18]. Even though the exact relation of spatial conformation of genes and their functionality is not fully understood yet, studies have shown that the structure of the human genome follows its functionality [16]. It is widely believed that chromosomal folding bring functional elements in close proximity regardless of their inter- or intra-chromosomal distance in base pair unit. In other words the concept of close and far in relation to the spatial map of genome is represented differently. Also, it has been shown that the contacts between small and gene-rich chromosomes are more frequent [18]. These evidences suggest linkages between chromosomal conformation, gene activity and their products (proteins) functionality. Identifying the significance of association of genomic location of genes and their products partner selection will aid in completion of the proximity pattern followed by genes and lead by their arrangement. The prospective pattern can be employed to the prediction models in order to infer potential protein interactions.

In this work we have selected the boundary of ideogram bands on each chromosome as genomic location of each gene. Several genes may reside on one ideogram as well as genes that may fall between two consecutive ideograms. The reason for this selection of distance unit is to take into account the effect of the co-expression of neighboring genes and the possible shared mechanism. Protein pairs from PPI dataset are mapped to their genomic location and simple

frequency calculation from retrieved data can identify the significance of interactions between two genomic segments. Based on these findings researchers can propose the genomic location pattern in which proteins preferentially select their interacting partners. These regions may contain genes involved in the same pathways or share the same functionality which are yet to be identified by further gene enrichment analysis. As shown in the provided SPARQL Query, we could retrieve the pre-determined inferred score between two ideograms, e.g. *Chromosome 3 - q29* and *Chromosome 10 - q24.32* containing the protein-coding genes for HES1 and SIRT1 (*entrezgene:23411*) proteins respectively.



(a) (b) (c)

Fig. 5. Illustration of the three use cases. **a) Use Case 1:** *Px* is the protein under query which consist of domain *x*. **b) Use Case 2:** *Dx* is the domain of interest independent of the containing protein. The returning result is a list of binary interactions between domain pairs which labels the interaction either as *validated* or *inferred*, the former is retrieved from 3did database while the latter is deduced from PPI data. **c) Use Case 3:** The genomic location of *PA* and *PB*, two interacting proteins, are mapped to their ideogrammatic location on two different chromosomes.

4 Related Work

Jiang et al. developed a semantic web base framework which predicts targets of drug adverse effect based on the PPIs and gene functional classification algorithm [11]. Chem2Bio2RDF [4] integrates data sources from Bio2RDF[9] in order to study polypharmocology and multiple pathway inhibitors which also requires thorough understanding of underlying PPI network.

5 Conclusions

The incorporation of complementary datasets for the expansion of PPI networks is a useful approach to gain insight into biological processes and to discover novel PPIs which have not been documented in the current PPI databases. However, there is an inherent high level of heterogeneity at the schema and instance level of these data sources, due to lack of a common representation schema and format. Hence, we decided to apply Linked Data concepts in the integration, retrieval and visualisation of concealed information. The enormous amount of publicly available data and its dynamicity, in terms of regular updates, is currently a

[9] `http://bio2rdf.org`

rate-limiting step to our data-warehousing approach for centralised analysis. We have proposed a domain-specific model which can accommodate the needs in the field of PPI modelling. The use of a domain-specific model and an interactive graph-based exploration platform for search and aggregative visualisation makes our integration approach more intuitive for the actual users who deal with PPI predictions. We have also proposed a set of three user scenarios depicting how LinkedPPI framework could be used for the prediction of potential interactions between proteins, domains and genomic regions.

6 Future Work

The approach which has been presented in this work is used in extraction of valuable information with regard to PPI network, domain-domain interactions and selective genomic interactions. However the observations reported in the outcome of such data retrieval is raw and could be a valuable asset for simulations and prediction methods if further analysis is done. As part of the future work we intend to apply statistical analysis on significance of such observations in order to be able to develop a classifier algorithm which is able to predict interacting and non-interacting protein pairs.

Acknowledgements This work has been done under the Simulation Science program at the National University of Ireland, Galway. SimSci is funded by the Higher Education Authority under the program for Research in Third-level Institutions and co-funded under the European Regional Development fund.

References

1. Alberts, B.: The Cell as a Collection of Protein Machines: Preparing the Next Generation of Molecular Biologists. Cell 92(3), 291–294 (Feb 1998)
2. Ashburner, M., Ball, C.A., et al.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics 25(1), 25–29 (May 2000)
3. Bhardwaj, N., Lu, H.: Correlation between gene expression profiles and protein-protein interactions within and across genomes. Bioinformatics 21(11), 2730–2738
4. Bin Chen, Xiao Dong, D.J.H.W.Q.Z.Y.D., Wild, D.J.: Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC Bioinformatics 11 (2010)
5. Bleda, M., Tarraga, J., de Maria, A., Salavert, F., et al.: CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources. Nucleic acids research 40(W1), W609–W614 (2012)
6. Chatterjee P, Basu S, K.M.N.M.P.D.: PPI$_S$VM: prediction of protein-protein interactions using machine learning, domain-domain affinities and frequency tables
7. Chen, H., Yu, T., Chen, J.Y.: Semantic Web meets Integrative Biology: a survey. Briefings in Bioinformatics 14(1), 109–125 (Jan 2013)
8. Flicek, P., et al.: Ensembl 2012. Nucleic acids research p. gkr991 (2011)
9. Goel, R., Harsha, H.C., Pandey, A., Prasad, K.S.: Human Protein Reference Database and Human Proteinpedia as resources for phosphoproteome analysis. Molecular bioSystems 8(2), 453–463 (Feb 2012)

10. Grigoriev, A.: On the number of protein-protein interactions in the yeast proteome. Nucleic Acids Res 31, 4157–4161 (2003)
11. Guoqian Jiang, Chen Wang, Q.Z., Chute, C.G.: A Framework of Knowledge Integration and Discovery for Supporting Pharmacogenomics Target Predication of Adverse Drug Events: A Case Study of Drug-Induced Long QT Syndrome. AMIA Summits Transl Sci Proc p. 8892 (2013)
12. Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., et al.: The HUPO PSI's molecular interaction formata community standard for the representation of protein interaction data. Nature biotechnology 22(2), 177–183 (2004)
13. Jelier, R., et al.: Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. Bioinformatics 21(9), 2049–2058 (2005)
14. Kamdar, M.R., Iqbal, A., Saleem, M., Deus, H.F., Decker, S.: GenomeSnip: Fragmenting the Genomic Wheel to augment discovery in cancer research. In: Conference on Semantics in Healthcare and Life Sciences (CSHALS). ISCB (2014)
15. Kamdar, M.R., Zeginis, D., Hasnain, A., Decker, S., Deus, H.F.: ReVeaLD: A user-driven domain-specific interactive search platform for biomedical research. Journal of biomedical informatics 47, 112–130 (2014)
16. Kosak, S.T., Groudine, M.: Form follows function: the genomic organization of cellular differentiation. Genes Dev 18, 1371–1384 (2004)
17. Licata, L., Briganti, L., et al.: MINT, the molecular interaction database: 2012 update. Nucleic acids research 40(Database issue), D857–D861 (Jan 2012)
18. Lieberman-Aiden, E., van Berkum, N., Williams, L., Imakaev, M., et al.: Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 326(5950), 289–293 (2009)
19. Liu, Z.P., et al.: Inferring a protein interaction map of mycobacterium tuberculosis based on sequences and interologs. BMC Bioinformatics 13(Suppl 7), S6 (2012)
20. Maali, F., Cyganiak, R., Peristeras, V.: Re-using cool uris: Entity reconciliation against lod hubs. In: Bizer, C., Heath, T., Berners-Lee, T., Hausenblas, M. (eds.) LDOW. CEUR Workshop Proceedings, vol. 813. CEUR-WS.org (2011)
21. Maglott, D., Ostell, J., Pruitt, K.D., Tatusova, T.: Entrez Gene: gene-centered information at NCBI. Nucleic acids research 39(Database issue), D52–7 (Jan 2011)
22. Michalak, P.: Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomic 91, 243248 (2007)
23. Obayashi, T., Okamura, Y., Ito, S., Tadaka, S., Motoike, I.N., Kinoshita, K.: COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals. Nucleic Acids Research 41(D1), D1014–D1020 (Jan 2013)
24. Ruepp, A., et al.: CORUM: the comprehensive resource of mammalian protein complexes - 2009. Nucleic Acids Research 38(Database-Issue), 497–501 (2010)
25. Ruttenberg, A., Clark, T., Bug, W., et al.: Advancing translational research with the Semantic Web. BMC bioinformatics 8 Suppl 3(Suppl 3), S2+ (2007)
26. Seal, R.L., Gordon, S.M., Lush, M.J., Wright, M.W., Bruford, E.A.: genenames. org: the HGNC resources in 2011. Nucl. Acids Res. 39(Suppl 1), D514–D519 (2011)
27. Sonnhammer, E.L., Eddy, S.R., et al.: Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28(3), 405–420 (Jul 1997)
28. Stark, C., Breitkreutz, B.J., Reguly, T., et al.: BioGRID: a general repository for interaction datasets. Nucleic acids research 34(suppl 1), D535–D539 (2006)
29. Stein, A., Russell, R.B., Aloy, P.: 3did: interacting protein domains of known three-dimensional structure. Nucleic acids research 33(suppl 1), D413–D417 (2005)
30. Uniprot-Consortium: The Universal Protein Resource (UniProt) 2009. Nucleic acids research 37(Database issue), D169–174 (Jan 2009)

Using the Micropublications ontology and the Open Annotation Data Model to represent evidence within a drug-drug interaction knowledge base

Jodi Schneider[1], Paolo Ciccarese[2], Tim Clark[2], and Richard D. Boyce[3]

[1] INRIA Sophia Antipolis France
jodi.schneider@inria.fr
[2] Massachusetts General Hospital and Harvard Medical School
paolo.ciccarese@gmail.com; tim_clark@harvard.edu
[3] University of Pittsburgh
rdb20@pitt.edu

Abstract. Semantic web technologies can support the rapid and transparent validation of scientific claims by interconnecting the assumptions and evidence used to support or challenge assertions. One important application domain is medication safety, where more efficient acquisition, representation, and synthesis of evidence about potential drug-drug interactions is needed. Potential drug-drug interactions (PDDIs), defined as two or more drugs for which an interaction is known to be possible, are a significant source of preventable drug-related harm. The combination of poor quality evidence on PDDIs, and a general lack of PDDI knowledge by prescribers, results in many thousands of preventable medication errors each year. While many sources of PDDI evidence exist to help improve prescriber knowledge, they are not concordant in their coverage, accuracy, and agreement. The goal of this project is to research and develop core components of a new model that supports more efficient acquisition, representation, and synthesis of evidence about potential drug-drug interactions. Two Semantic Web models—the Micropublications Ontology and the Open Annotation Data Model—have great potential to provide linkages from PDDI assertions to their supporting evidence: statements in source documents that mention data, materials, and methods. In this paper, we describe the context and goals of our work, propose competency questions for a dynamic PDDI evidence base, outline our new knowledge representation model for PDDIs, and discuss the challenges and potential of our approach.

Keywords: Linked Data, drug-drug interactions, evidence bases, Micropublications, Open Annotation Data Model, knowledge bases

1 Introduction

Scientific knowledge depends on the verification and integration of large systems of interconnected assertions, assumptions, and evidence. These systems are con-

tinually growing and changing, as new scientific studies are completed and new documents are published. The state of current knowledge in any given domain can be difficult for any one individual to fully grasp, because bits of knowledge are updated at frequent intervals.

In the biosciences, this problem has taken on particular importance, due to an exponential growth in the aggregate publication rate. Manually curated databases are used to record certain types of knowledge. To update and maintain these databases, curators must make knowledge-intensive decisions, identifying the best available evidence in the current scientific literature. Maintaining such databases is challenging because there is limited tracking of the source information.

In an ongoing project, we are experimenting with using the Micropublications Ontology[4] [Clark2014] and the Open Annotation Data Model[5] [W3C2013] to create an audit trail between assertions, evidence, and source documents, so that assertions and evidence can be flagged for update in flexible and intelligent ways. Updates may be needed when the underlying sources change, when a particular method for establishing an assertion is discredited, etc. Our goal is to provide better linkages between an assertion recorded in a knowledge base and its supporting evidence (i.e., data, materials, and methods) found in source documents.

In the remainder of the paper, we describe the competency questions for our evidence base and the new evidence model that we are creating, which combines the Micropublication Ontology and the Open Annotation Data Model, and adapts them to the existing evidence modeling of the Drug Interaction Knowledge Base[6] [Boyce2007,Boyce2009]. We then reflect on how the new model performs for our goal of creating an audit trail between assertions, evidence, and source documents.

2 Context and goals

Our work is in the context of a larger project on organizing and synthesizing scientific evidence from the biomedical literature on potential drug-drug interactions. Potential drug-drug interactions (PDDIs), defined as two or more drugs for which an interaction is known to be possible, are a significant source of preventable drug-related harm (i.e., adverse drug events, or ADEs). The combination of poor quality evidence on PDDIs, and a general lack of PDDI knowledge by prescribers, results in many thousands of preventable medication errors each year. While many sources of PDDI evidence exist to help improve prescriber knowledge, they are not concordant in their coverage [Saverno2011], accuracy [Wang2010], and agreement [Abarca2003]. Difficulties with synthesizing evidence, and gaps in the scientific knowledge of PDDI clinical relevance, underlie such disagreement.

[4] `http://purl.org/mp/`
[5] `http://www.openannotation.org/spec/core/`
[6] `http://purl.net/net/drug-interaction-knowledge-base/`

To address these problems, our research group is studying the potential benefit of applying recent developments from the Semantic Web community on scientific discourse modeling and open annotation. The goal is to develop core components of a new PDDI knowledge representation model that will support a more efficient acquisition, representation, and synthesis of PDDI evidence. The desired knowledge representation will provide better linkages between PDDI assertions and their supporting evidence, by directly connecting to annotated section(s) of relevant source documents.

3 Approach

Our new approach will draw upon the current version (1.2) of the Drug Interaction Knowledge Base [Boyce2007,Boyce2009], the Open Annotation Data Model [W3C2013], and the Micropublications Ontology [Clark2014].

The Drug Interaction Knowledge Base (DIKB) is a static, manually constructed evidence base that indexes assertions and evidence of PDDI for over 60 drugs. Its taxonomy of assertion types and evidence types [Boyce2014] is a starting point for the new knowledge base. The current version of the DIKB implements a version of the SWAN semantic discourse ontology [Ciccarese2008] to represent evidence relations. Specifically, the knowledge base uses *swanco:citesAsSupportingEvidence* and *swanco:citesAsRefutingEvidence* to link to an entire source document as a supporting or refuting citation. At the time the DIKB 1.2 was constructed (2007–2009), annotation methodologies were less well developed. Consequently, version 1.2 of the DIKB stores quotes as textual strings manually copied from source documents. The text has been enriched with metadata about the source section, but it is non-trivial to return to the appropriate segment of the text from this information.

Our use of the Open Annotation Data Model (OA) reflects a change in the state of the art. OA is an "an interoperable framework for creating associations between related resources, annotations, using a methodology that conforms to the Architecture of the World Wide Web"[7]. In particular, OA allows an evidence database to provide explicit connections from quotes to their source documents. For example, as shown in Figure 1, an OA resource can be used to quote a specific part of a drug product label (also known as a summary of product characteristics) to indicate evidence that *escitalopram inhibits CYP2D6*. In general, OA enables queryable links between selections from source documents (as target) to the instances of data, methods, and materials (as body) that we want to model to support drug interaction knowledge base use cases.

Similarly, the Micropublications Ontology improves the depth with which evidence can be represented and queried. The most important feature of the Micropublications model, in our view, is its ability to represent the data, methods, and materials that act as support for a claim, and to transitively close chains

[7] `http://www.openannotation.org/spec/core/`

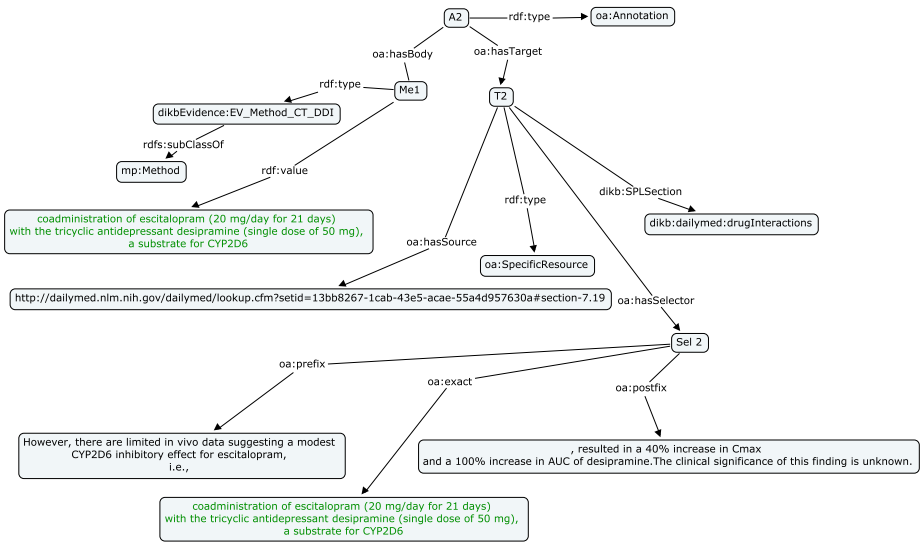


Fig. 1. The Open Annotation ontology (*oa*) can be used to quote the evidence (here a method described in a DailyMed product label) and associate it with an instance of the Micropublication ontology (*mp*). The annotation records quoted text (*oa:exact*) as the target, while the body of the annotation is a *mp:Method* instance, *Me1*, supporting the Claim *Escitalopram inhibits CYP2D6* shown in Figure 2. We use existing terminology from the DIKB ontology to specify the section of the DailyMed product label and to indicate the *dikbEvidence* type.

of claims[8] and citations across the literature to their fundamental supporting evidence. A *mp:Micropublication mp:argues* a *mp:Claim* based on connecting any number of *mp:Representation*s. The whole Micropublication is a Representation, as are Data and Methods (including Materials and Procedures), whether textual or pictoral. A *mp:Representation* may *mp:support* or *mp:challenge* any other *mp:Representation*, making the evidence explicit and queryable.

4 Competency Questions

To design an appropriate enhancement of the DIKB model with Micropublications and the Annotation Ontology, we need to understand what sorts of questions experts would like to retrieve about the PDDIs. The competency questions below were elicited from experienced editors of clinically oriented drug compendia during the process of developing DIKB 1.2. Most fall into three categories: finding assertions and evidence; assessing the evidence; and enabling updates. A second area of interest is statistical information about the evidence base which is useful for various analytics related to knowledge base maintainance.

4.1 Finding assertions and evidence

1. **Finding assertions:**
   (a) List all assertions that are not supported by evidence
   (b) Which assertions are supported (or refuted) by just one type of evidence?
   (c) Which assertions have evidence from source X (e.g., product labeling)
   (d) Which assertions have both evidence for and evidence against from a single source X?
2. **Finding evidence:**
   (a) List all evidence for or against assertion X (by evidence type, drug, drug pair, transporter, metabolic enzyme, etc.)
   (b) What is the in vitro evidence for assertion X? the in vivo evidence?
   (c) List all evidence that has been flagged as rejected from entry into the the knowledge base
   (d) Which single evidence items act as support or rebuttal for multiple assertions of type X (e.g., $substrate\_of$ assertions)?

4.2 Assessing the evidence:

1. **Understanding evidence coming from a given study:**
   (a) What data, methods, materials, are reported in evidence item X?
   (b) Which evidence items are related to and follow-up on evidence item X?
   (c) Which research group conducted the study used for evidence item X?
   (d) Are the evidence use assumptions for evidence item X concordant? unique? non-ambiguous?

[8] 'Assertion' in DIKB terminology corresponds to a 'Claim' in the Micropublications model; this variation in terms is because the term 'claim' is used in a different sense in medical billing.

2. **Verifying plausibility of an evidence item:**
   (a) Has evidence item X been rejected for assertion Y? If so, why and by whom?
   (b) Which other assertions are being supported/challenged by this evidence item?
   (c) What are the assumptions required for use of this evidence item to support/refute assertion X?
3. **Checking assertions about pharmacokinetic parameters (i.e., area under the concentration time curve (AUC))**
   (a) How many pharmacokinetic studies used for evidence items in the DIKB could be used to support or refute an assertion about pharmacokinetic paramater X (e.g., 'X increases AUC')?
   (b) How many pharmacokinetic studies in the DIKB used for evidence items for assertion X are based on data from the product label?
   (c) What is the result of averaging (or applying some other statistical operation) to the values for pharmacokinetic parameter X across all relevant studies used for evidence items?
4. **Checking for differences in the product labeling:**
   (a) Are there differences in the evidence items that were identified across different versions of product labeling for the same drug?
   (b) What version of product labeling was used for evidence item X? Original manufacturer or repackager? Most current label or outdated? Is the drug on market in country X or not? American or country X?

4.3 Supporting updates to evidence and assertions

1. **Changing status of redundant and refuted evidence:**
   (a) Remove a older version of a redundant evidence item
   (b) Change the modality of a supporting evidence item to be a refuting evidence item
2. **Updating when key sources change:**
   (a) Get all assertions that are supported by evidence items identified from an FDA guidance or other source document just released as an updated version.

4.4 Understanding the evidence base

1. **Statistical information about the evidence base:**
   (a) Number of assertions in the system
   (b) Number of evidence items for and against each assertion type
   (c) Show the distribution of the levels of evidence for various assertion types (e.g., pharmacokinetic assertions)

5 Modeling evidence about drug-drug interactions

Figure 2 shows how the new Micropublications model of evidence on PDDIs would represent some of the evidence supporting and challenging the assertion *escitalopram does not inhibit CYP2D6*. We created the example by hand using a sample assertion and evidence items from the DIKB version 1.2[9].

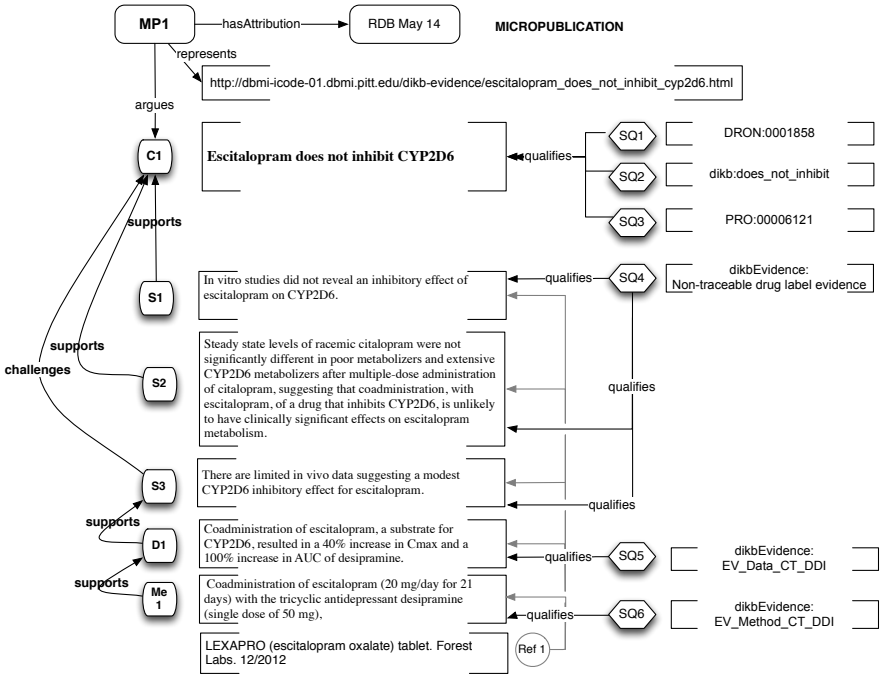


Fig. 2. A model of the evidence for and against the assertion *escitalopram does not inhibit CYP2D6*. This is based on the Micropublications ontology, and reuses the evidence taxonomy (dikbEvidence), terms (dikb), and data from the DIKB. The Drug Ontology (DRON) and Protein Ontology (PRO) are reused in semantic qualifiers. A more detailed view of Method *Me1* is shown in Figure 1.

The Micropublications ontology is used to structure the evidence relating to data, methods, and materials, and the overall indication that evidence *mp:supports* or *mp:challenges* a *mp:Claim*. We qualify Claims (C1 in the figure) by reusing identifiers from DRON[10] [Hanna2013] and the Protein Ontology[11] [Natale2011].

[9] `http://dbmi-icode-01.dbmi.pitt.edu/dikb-evidence/escitalopram_does_not_inhibit_cyp2d6.html`

[10] `http://purl.obolibrary.org/obo/dron.owl`

[11] `http://pir.georgetown.edu/pro/`

The new model reuses the DIKB evidence taxonomy[12] to provide epistemic qualification (SQ2, SQ5, SQ6 in the figure) to statements (S1, S2, and S3 in the figure), data (D1 in the figure), methods (Me1 in the figure), and materials (not shown in this example). The Open Annotation Data Model (previously shown in Figure 1) is used to link quotes taken from source documents back to their originating information artifacts. The approach to modeling other DIKB assertions would be similar to this example.

6 Discussion

6.1 Expected Benefits

Certain benefits accrue from upgrading from the current DIKB. Many of the competency questions (Section 4) are not supported in the DIKB 1.2. The new model is designed to support these and additional questions relevant in the domain. Visual inspection of the model suggests that we will be able to answer some competency questions quite naturally. In particular, finding the assertions that are not supported by evidence already in the evidence base, the evidence that should be checked most thoroughly (e.g. evidence that by itself supports multiple assertions), and the data, methods, and materials associated with a given evidence item as described in source documents.

Further, as a Linked Data resource, our new knowledge base will also enable innovative queries using knowledge from other sources about tagged entities (i.e., drugs and proteins) represented in the evidence base. Unlike the current DIKB, we will be able to render annotations in their original context. We also expect to be able to support distributed community annotation/curation, since MP and OA take account of provenance, and since OA is being increasingly adopted by a variety of annotation tools.

6.2 Modeling challenges

Our project does raise certain modeling challenges. To date, MP has not been used to represent both unstructured claims and the related logical sentences. Figure 1 shows the assertion *escitalopram does not inhibit CYP2D6* as unstructured text. However, the DIKB requires that 1) assertions about PDDIs be formulated by experts prior to collecting evidence, and 2) that the assertions be represented both as unstructured statements and sentences in a logical formalism. Careful thought is being put into how to properly accommodate this use case. Such challenges are to be expected since MP is a relatively new ontology and since this is a new application of it.

Another challenge is to ensure that, as the evidence base scales, competency questions can be answered efficiently. To address this, we building the model using an iterative design-and-test approach. In this process, efficient querying is a key requirement.

[12] `http://bioportal.bioontology.org/ontologies/DIKB`

6.3 Other issues

For enabling synthesis over the PDDI information, the model is not the only concern. Applying this model will require integration work. One challenge is inherent to scholarly documents: the existing evidence items within the DIKB refer to many data, materials, and methods that exist only in PDF documents accessible only through proprietary portals or academic library systems. Consequently, resolving annotations requires a method for pointing to proprietary *oa:target*s.

7 Conclusions & Future Work

We are currently iterating and refining the PDDI evidence and annotation model. Once it is stable, we plan to use the new model to represent as Linked Data evidence collected by an evidence board consisting of drug experts. The evidence collection effort is planned as part of a research project funded by the National Library of Medicine ("Addressing gaps in clinically useful evidence on drug-drug interactions", 1R01LM011838-01) and will focus on PDDI assertions for a number of commonly prescribed drugs (anticoagulants, statins, and psychotropics). We plan to implement a pipeline for extracting PDDI mentions from a variety of publicly available sources, including published journal articles indexed in PubMed or PubMed Central, FDA Guidance Documents, and drug product labels from the National Library of Medicine's DailyMed website[13]. Candidate PDDI assertions will be linked by machine to the Internet-accessible versions of the information artifacts used as evidence.

An existing Micropublication plugin for Domeo [Ciccarese2014] is being modified as part of the project. Our plan is to use the revised plugin to support the evidence board with the collection of the evidence and associated annotation data. It will also enable the broader community to access and view annotations of PDDIs highlighted in a web-based interface. We anticipate that this approach will enable a broader community of experts to review each PDDI recorded in the DIKB and examine the underlying research study to confirm its appropriateness and relevance to the evidence base.

The usability of the annotation plug-in is critically important so that the panel of domain experts will not face barriers to annotating and entering evidence. This will require usability studies of the new PDDI Micropublication plugin. Another issue is that many PDDI evidence items can be found only in PDF documents. Currently, the tool chain for PDF annotation is relatively weak: compared to text and HTML, PDF annotation tools are not as widely available and not as familiar to end-users. Suitable tools will have to be integrated into the revised plugin.

Knowledge representations combining MP and OA have the potential to allow more granular and reusable representation of evidence (data, materials, and methods), which are needed for synthesizing contested knowledge at the state

[13] `http://dailymed.nlm.nih.gov/dailymed/about.cfm`

of the art from scientific documents. The knowledge representations we are now creating will be beneficial for integrating PDDI evidence, and we hope they will inspire an increased use of linked data for evidence synthesis in other domains.

Acknowledgments

This work was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 246016, and a grant from the National Library of Medicine (1R01LM011838-01). We thank Carol Collins, Lisa Hines, and John R Horn for serving on the Evidence Panel of "Addressing PDDI Evidence Gaps", and for contributing to the competency questions presented here.

References

[Abarca2003] Abarca, Jacob, Daniel C. Malone, Edward P. Armstrong, Amy J. Grizzle, Philip D. Hansten, Robin C. Van Bergen, and Richard B. Lipton. "Concordance of severity ratings provided in four drug interaction compendia." *Journal of the American Pharmacists Association* 44;2 (2003): 136–141.

[Boyce2014] Boyce, R.D. "A Draft Evidence Taxonomy and Inclusion Criteria for the Drug Interaction Knowledge Base." August 9, 2014, url: http://purl.net/net/drug-interaction-knowledge-base/evidence-types-and-inclusion-criteria

[Boyce2007] Boyce, Richard D., Carol Collins, John Horn, and Ira Kalet. "Modeling Drug Mechanism Knowledge Using Evidence and Truth Maintenance." *IEEE Transactions on Information Technology in Biomedicine* 11;4 (2007): 386–397.

[Boyce2009] Boyce, Richard D., Carol Collins, John Horn, and Ira Kalet. "Computing with evidence: Part I: A drug-mechanism evidence taxonomy oriented toward confidence assignment." *Journal of Biomedical Informatics* 42;6 (2009): 979–989.

[Ciccarese2008] Ciccarese, Paolo N., Elizabeth Wu, Gwen Wong, Marco Ocana, June Kinoshita, Alan Ruttenberg, and Tim Clark. "The SWAN biomedical discourse ontology." *Journal of Biomedical Informatics* 41;5 (2008): 739–751.

[Ciccarese2014] Ciccarese, Paolo N., Marco Ocana, and Tim Clark. "Open semantic annotation of scientific publications using DOMEO." *Journal of Biomedical Semantics* Apr 24;3 (2012): Suppl 1:S1.

[Clark2014] Clark, Tim, Paolo N. Ciccarese, and Carole A. Goble. "Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications." *Journal of Biomedical Semantics* 5;28 (2014).

[Hanna2013] Hanna, Josh, Eric Joseph, Mathias Brochhausen, and William R. Hogan. "Building a drug ontology based on RxNorm and other sources." *Journal of Biomedical Semantics* 4 (2013): 44–52.

[Natale2011] Natale, Darren A., Cecilia N. Arighi, Winona C. Barker, Judith A. Blake, Carol J. Bult, Michael Caudy, Harold J. Drabkin, Peter D'Eustachio, Alexei V. Evsikov, Hongzhan Huang, Jules Nchoutmboube, Natalia V. Roberts, Barry Smith, Jian Zhang and Cathy H. Wu. "The Protein Ontology: a structured representation of protein forms and complexes." *Nucleic acids research* 39, no. suppl 1 (2011): D539–D545.

[Saverno2011] Saverno, Kim R., Lisa E. Hines, Terri L. Warholak, Amy J. Grizzle, Lauren Babits, Courtney Clark, Ann M. Taylor, and Daniel C. Malone. "Ability of pharmacy clinical decision-support software to alert users about clinically important drug-drug interactions." *Journal of the American Medical Informatics Association* 18;1 (2011): 32–37.

[Wang2010] Wang, Lorraine M., Maple Wong, James M. Lightwood, and Christine M. Cheng. "Black box warning contraindicated comedications: concordance among three major drug interaction screening programs." *Annals of Pharmacotherapy* 44; 1 (2010): 28–34.

[W3C2013] Sanderson, Rob, Paolo N. Ciccarese, and Herbert Van de Sompel (editors). "Open Annotation Data Model", W3C Community Group Draft, 08 February 2013, url: http://www.openannotation.org/spec/core/

Capturing Provenance for a Linkset of Convenience

Simon Jupp[1], James Malone[1], and Alasdair J G Gray[2]

[1] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, United Kingdom
[2] Department of Computer Science, Heriot-Watt University, Edinburgh, United Kingdom

Abstract. Biological interactions such as those between genes and proteins are complex and require intricate OWL models. However, direct links between biological entities can support search and data integration. In this paper we introduce *linksets of convenience* that capture these direct links. We show the provenance statements required to track the derivation of such linksets; linking them back to the full biological justification.

Keywords: Data linking, Provenance, VoID

1 Introduction

Investigating biological systems, such as those implicated in disease, necessitates the connection of many levels of biology; gene, gene variation, gene expression, protein structure, signalling pathways, phenotypic, epidemiological data and so on. The ability to integrate data across these levels relies on links that can be formed between biological entities, for example, going from a gene to proteins or proteins to pathways. For each of these links there is some biological justification that may involve several steps (see Section 2 for details). To support tasks such as search and data integration it is convenient to provide additional shortcuts in the form of a direct link, e.g. genes to pathways.

Modeling the true nature of the links using semantic web technologies such as OWL removes ambiguity when working with data by giving it a well defined and precise semantics. However it increases the complexity of interacting with the data as the OWL model needs to capture the full intricacies of the biological interactions. As we move to publish biological data as linked open data, there is an opportunity to describe direct links between different types of biological entities as a shortcut to be made between entities which feature in common queries, such as gene to protein; capturing the way that biologists often discuss the domain and enable novel integrations of the data. These direct links provide a working notion that cuts through the biology but which does not necessitate capturing (or recapturing) the complex multivariate relationships that can hold between the two entities. Such linksets are already used to support the Open

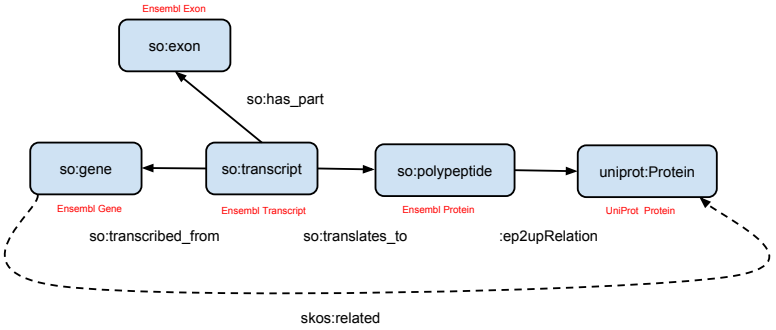


Fig. 1. Linking an Ensembl gene with its UniProt protein. Solid lines show the full semantic modelling required while the dashed line represents the linkset of convenience.

PHACTS Discovery Platform [1], although those linksets do not have adequate provenance.

In this paper we propose a mechanism to model these *links of convenience* using a combination of VoID linksets [2] and PROV [3]. We avoid misrepresenting links by applying semantically weaker relationships together with additional provenance which represents the underlying complexity. We illustrate the model with an example using data from two popular biological databases.

2 Linking genes to proteins use case.

We motivate our work with an example mapping between Ensembl [4] (a database of genome annotation) and Uniprot [5] (a database of protein sequences). These databases already contain cross-references between an Ensembl Gene (EG) and a Uniprot Protein (UP). However to understand how this mapping is generated you currently need to discover the correct publications and online documentation; they are not directly discoverable from the data.

Biological theory tells us that a gene encodes for a protein, although this biological relation only truly holds for the link between the EG and the Ensembl Protein (EP) entity. There are in fact multiple types of UP to EP mappings, for instance they can be derived from an exact sequence identity or they might be based on a percentage sequence identity. Figure 1 illustrates how we model EG to EP using terminology defined in the Sequence Ontology, and for illustration we include a superproperty of the all the EP to UP mappings that we call `ep2upRelation`[3]. We introduce a link of convenience (dashed line) that links the EG to UP that is there to support queries using the semantically weak `skos:related` relation. This schema lacks the provenance to assert that the related link of convenience is derived from the longer chain of semantically richer links that hold from a gene to protein.

[3] UniProt are currently extending their vocabulary to define these relations.

```
# define the ensembl protein partition
:ensembl void:classPartition :EPpartition .
:EPpartition void:class so:Polypeptide .

# define the Uniprot protein partition
:uniprot void:classPartition :UPpartition .
:UPpartition void:class uniprot:Protein .

# define the linkset that links the two partitions
:ensemblProteinToUniprotProteinLinkset a void:Linkset ;
    void:linkPredicate :ep2upRelation ;

# define partitions for ensembl gene, gene transcript and
# transcript protein
:ensembl  void:classPartition :ensemblGenePartition ;
    void:propertyPartition :ensemblGeneTranscriptPartition ;
    void:propertyPartition :ensemblTranscriptProteinPartition ;
:ensemblGenePartition void:class so:gene .
:ensemblGeneTranscriptPartition void:property so:transcribed_from .
:ensemblTranscriptProteinPartition void:property so:translates_to .

# define the linkset that links the two partitions,
# including the dataset description that contains the triples that
# are used to derive this linkset
:ensemblGeneToUniprotProteinLinkset a void:Linkset ;
    void:linkPredicate skos:related ;
    void:subjectsTarget :ensemblGenePartition;
    void:objectsTarget :UPpartition;
    prov:wasDerivedFrom :ensemblGeneTranscriptPartition,
        :ensemblTranscriptProteinPartition,
        :ensemblProteinToUniprotProteinLinkset
```

Fig. 2. Description of the linkset of convenience between Ensembl Gene and UniProt Protein which includes the provenance derivation.

3 Describing Linksets

The model outlined in Figure 1 can be decorated with provenance that captures additional information about how the link of convenience between EG and UP is derived. The resulting linkset description is shown in Figure 2. In the following we describe the blocks of RDF.

The VoID vocabulary of linked datasets allows the description of RDF links between datasets using VoID linksets. A linkset allows us to describe the links, captured as a set of triples, between two datasets. We can use VoID to describe relevant partitions of the datasets based on individual properties or classes, these form new subsets that can participate in multiple linksets. In our scenario we

need to capture two crucial linksets; the first is the EP to UP linkset, and the second is the more convenient EG to UP linkset.

The EP-UP linkset captures the `:ep2upRelation` link between types of EP in the Ensembl dataset, and types of UP in the UniProt dataset (lines 10-11). We describe two further subsets; the EP partition of all entities that are of type `so:Polypeptide` in the Ensembl dataset (lines 2-3) and the UniProt subset of all entities that are of type `uniprot:Protein` (lines 6-7).

The EG to UP link of convenience needs a similar linkset description based on an EG partition and the previous UP partition, although this time the relation is `skos:related` (lines 25-26). We also want to capture that the triples in this linkset are derived from another set of triples. This captures that the `skos:related` is a shortcut relation for a more complex path through the RDF graph. Again we can use VoID partitioning, but this time using a property based partition to identify the EG to Ensembl Transcript (ET) and ET to EP links (lines 15-20) . Finally we use the `prov:wasDerivedFrom` relation to link the convenience linkset to the linksets that describe the full path of relations that the shortcut represents (line 28-30).

4 Discusion

It is always important to try and model your data as accurately as possible, and publishing data with RDF and OWL is well suited for this task. The VoID vocabulary already provides a mechanism to define and attach provenance to linksets between datasets, and we are proposing the use of PROV to connect linksets that are derived from other linksets. As a Web of linked biological data emerges, there is a need to identify links that are there for convenience, and expose how they relate back to the core biological (OWL) model. In cases where a link of convenience is derived from a series of other linksets, it is desirable to be able to spot this and unpack the convenience links using common queries. The model proposed supports this task but questions remain as to whether VoID and PROV are enough, so we hope this preliminary work can help motivtate the discussion.

Acknowledgements

EBI contribution supported by EU FP7 BioMedBridges Grant 284209.

References

1. Gray, A.J.G., Groth, P., Loizou, A., Askjaer, S., Brenninkmeijer, C.Y.A., Burger, K., Chichester, C., Evelo, C.T., Goble, C.A., Harland, L., Pettifer, S., Thompson, M., Waagmeester, A., Williams, A.J.: Applying linked data approaches to pharmacology: Architectural decisions and implementation. Semant. Web **5** (2014) 101–113
2. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing Linked Datasets with the VoID Vocabulary. Note, W3C (March 2011)

3. Lebo, T., Sahoo, S.S., Mcguinness, D.: PROV-O: The PROV Ontology. Technical report, W3C Recommendation (2013) http://www.w3.org/TR/prov-o/.
4. Flicek, P., Amode, M.R., Barrell, D., et al: Ensembl 2014. Nucleic acids research **42** (2014) D749–D755 doi: 10.1093/nar/gkt1196.
5. The UniProt Consortium: Activities at the universal protein resource (UniProt). Nucleic acids research **42** (2014) D191–D198 doi: 10.1093/nar/gkt1140.

Connecting Science Data Using Semantics and Information Extraction

Evan W. Patton and Deborah L. McGuinness

Rensselaer Polytechnic Institute
110 8^{th} Street, Troy, NY 12180 USA
{pattoe, dlm}@cs.rpi.edu

Abstract. We are developing prototypes that explicate our vision of connecting personal medical data to scientific literature as well as to emerging grey literature (e.g., community forums) to help people find and understand information relevant to complex medical journeys. We focus on robust combinations of natural language processing along with linked data and knowledge representation to build knowledge graphs that help people make sense of current conditions and enable new manners of scientific hypothesis generation. We present our work in the context of a breast cancer use case. We discuss the benefits of biomedical linked data resources and describe some potential assistive technology for navigating rich, diverse medical content.

Keywords: knowledge representation, explanation, clinical notes, natural language, web forums, nanopublications

1 Introduction

As scientific knowledge continues to grow in size and diversity, it is increasingly difficult to discover and manage information relevant to any particular context. It can be challenging to determine how a statement or report relates to others and to form and evaluate (often competing) hypotheses, e.g. related to diagnosis or treatment paths. Complications grow when content is both structured and unstructured, and when some is from less accredited sources. We aim to expand the boundaries of Linked Science by focusing on evidence modeling from natural language processing techniques (NLP) over broad content and by identifying promising data-driven hypotheses using linked data and nanopublication style encodings. We present this discussion in the context of a breast cancer demonstration use case informed by challenges experienced during a co-author's recent cancer journey. Cancer is a complex disease to manage and treat, often requiring chemotherapy, surgery, radiation, and drugs to reduce recurrence. We show how management of this information by the patient is aided by semantic technologies combined with natural language processing algorithms.

> A breast cancer patient wishes to better understand her diagnosis and planned treatment. She is interested in expected chemotherapy side effects, and leveraging experiences of other similar individuals to proactively find and evaluate promising coping strategies. She reads through

> oncologist-provided documents about her proposed chemotherapy drugs and uses search engines to find more about likely adverse effects that appear detrimental to her quality of life. She finds conflicting opinions on the efficacy of different coping strategies, and needs to determine an approach to effectively weigh the possible pros and cons. Managing this information is mentally taxing and can easily overwhelm a patient.

Our patient needs to find and comprehend potentially conflicting evidence about treatment options and side effects. We propose new software, using a variety of artificial intelligence tools built on the interoperability principles promulgated by linked data and the Semantic Web, to address these challenges.

2 Evidence Modeling

The patient uses current technologies to obtain information about her treatment strategy and to formulate promising side effect mitigations. This can be time consuming for anyone, but more so for medically naïve patients. Furthermore, technologies such as web forums or social networking sites are becoming increasingly common for discourse between patients as they can often include anecdotal reports, that have not yet been validated through clinical trials, but may be valuable. They are often presented in layperson terms and sometimes attract new patients who may be less medically literate. Due to lack of scientific rigor, there may be contradictory or unsupported information available, as shown in the following two answers about a mitigation for the very common, taxol-related, nail bed problem:

> My onc[ology] nurse told me to rub tea tree oil into my cuticles and nails every night. It is a natural anti-septic and for whatever reason can sometimes help prevent nail infections and lifting during taxol. [1]

> I wouldn't use tea tree oil. A friend did on some cracked skin and it got worse. [2]

The first suggestion is a common preventive approach for nail problems: tea tree oil prevents nail infections because "it is a natural anti-septic" and appeals to authority "my onc nurse told me to...". The second suggestion from a different user in the same thread advises against tea tree oil as "a friend [applied tea tree oil] on some cracked skin and it got worse." Natural Language techniques may be used to extract coping strategies for particular conditions but without deeper knowledge, provenance, and tools, the user may not know how to evaluate and/or integrate potentially contradictory suggestions. We are extending joint extraction techiques proposed in [4] with semantic background knowledge to aid in extracting linked data from medical records.

[1] https://community.breastcancer.org/forum/69/topic/783573
[2] https://community.breastcancer.org/forum/96/topic/745475

3 Hypothesis generation using Nanopublications

The Repurposing Drugs using Semantics (ReDrugS) project [5] has focused on modeling evidence using small units of publishable information called Nanopublications [2]. ReDrugS utilizes linked data sources to build a knowledge base of nanopublications that is then reasoned about using probabilistic techniques to identify potential links between proteins, drugs, binding sites, and genes, with the ultimate aim of discovering possible new off-label uses for FDA-approved drugs. This project's success has been partially due to the large corpus of linked data and ontologies generated by the biomedical community over the past few decades. ReDrugS has ingested content from 17 structured curated data sources, including content concerning drugs, alternate names, conditions, and pathways. Once a chemotherapy protocol is extracted from medical notes, ReDrugs can be used to find alternative drug names along with related conditions. This framework, along with the side effect resource SIDER in process, can be used to improve the patient's process in finding chemotherapy drug side effects and some mitigations by applying its search techniques to authoritative drug resources, such as looking for anti-nausea prescription drugs. The infrastructure for this system could be repurposed for other scientific domains, but only if linked data sources are abundant in those domains or if quality linked data can be generated from automated methods, e.g. via natural language processing of web-based resources.

4 Explanations

We aim to provide extensive explanation mechanisms since explanation is a key component of transparent systems and user studies have shown that explanations are required if agents are to be trusted [1]. We aid explanation generation through the collection of provenance, modeled using the W3C's PROV ontology [3]. PROV-O is a standard for modeling provenance information on the web, which allows tools to integrate distributed provenance information from different systems. We use this provenance to help construct end user explanations that include both lineage of content and support (and opposition) for a statement.

We identify potential evidence on the use of tea tree oil in chemotherapy-induced nail bed problems. Not only would a patient want to know evidence, source, and authoritativeness for both views, she might also want the system further decompose these arguments and present supporting evidence as to the antimicrobial nature of tea tree oil in more authoritative sources (e.g. [6]).

We claim that we can reuse the ReDrugS content to find prescription drugs for chemotherapy side effects. Provenance may be displayed to show that the recommendation is from a validated authoritative source. While that framework was originally designed to find potential new off-label uses for drugs along with confidence ratings, the explanation component is more critical for our use so that researchers may inspect evidence sources and the methods used to determine the system confidence. Without such explanations, people would have difficulty evaluating competing suggestions.

Our systems[3] provide explanation drill down so users can obtain as much detail as they desire, thus allowing a patient to find, for example, if authoritative sources contain prescription drugs for coping with a particular side effect. Our NL-based extraction work can be used to identify alternative, possibly competing, therapies, e.g. an herbal remedy recommended anecdotally with potentially corroborating authoritative sources.

5 Discussion and Summary

Natural Language Processing can expose some of the unstructured content of medical records as structured content as well as assist in generating linked data from unstructured sources. The ReDrugS framework provides a semantically-integrated system combining many different structured biomedical resources to generate a broadly reusable knowledge graph. By integrating the natural language and structured knowledge representation approaches, we can obtain a much richer annotated knowledge base that includes source and confidence information. Our prototypes demonstrate some ways that this rich resource may then be used to help patients and their support networks to discover, integrate, and evaluate information relevant to complicated medical situations and to help form transparent and data-driven hypotheses about how to proceed. We believe these efforts demonstrate some opportunities for future AI-enhanced Linked Science-based assistants that use the wealth of structured content as well as the growing grey literature collection.

Acknowledgements

The authors thank Heng Ji and Alex Borgida for their discussions that helped shape this work.

References

1. Glass, A., McGuinness, D.L., Wolverton, M.: Toward establishing trust in adaptive agents. In: 13th Intl Conference on Intelligent User Interfaces. pp. 227–236 (2008)
2. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Information Services & Use 30, 51–56 (2010)
3. Lebo, T., Sahoo, S., McGuinness, D.L.: PROV-O: The PROV ontology. Tech. rep., W3C (2013)
4. Li, Q., Ji, H.: Incremental joint extraction of entity mentions and relations. In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (2014)
5. McCusker, J., Solanki, K., Chang, C., Dumontier, M., Dordick, J., McGuinness, D.L.: A nanopublication framework for systems biology and drug repurposing. In: CSHALS 2014 (2014)
6. Pazyar, N., Yaghoobi, R., Bagherani, N., Kaerouni, A.: A review of applications of tea tree oil in dermatology. International Journal of Dermatology pp. 784–90 (2013)

[3] http://tw.rpi.edu/web/project/MobileHealth
http://tw.rpi.edu/web/project/ReDrugS