Chapter 7
Conclusions

Abstract In this chapter, I conclude the discussion of my work on Semantic Publishing. In particular, I summarise my own personal contributions in order to address one of the main issues of this field, i.e., the linking of a text to the formal representation of its meaning and thus the representation of its structure and of its argumentative discourse. In addition, I summarise my own contribution on the development of interfaces to hide the complexity of markup and ontology formalisms behind user-friendly views in order to help users of Semantic Publishing (e.g., scholars, publishers, archivists, librarians, etc.) that may have difficulties in interacting with Semantic Publishing technologies. Finally, I conclude the chapter introducing planned future works for all the languages, models and tools presented.

In the early days of the Web, the intrinsic meaning of the content of a document such as a Web page was accessible only to human readers, using their capabilities to conceptualise the particular semantics starting from natural language descriptions. The Semantic Web was born from a desire to develop mechanisms for machine understanding of that same content that would be as effective as that of humans. Its final goal was to "bring structure to the meaningful content of Web pages" and to provide "a new class of tools by which we can live, work and learn together" (Berners-Lee et al. 2001). In other words, it tried to link authored text (i.e., Web pages) to its formal semantics in a way that "intelligent" applications can be developed so as to significantly assist people in their everyday life.

To this end, the Semantic Web communities initially started to develop standards and technologies with the aim of giving a theoretical and practical background to enable the creation of intelligent applications and enhanced Web resources. Starting from these bases, recently some research and institutional domains are trying to make a further step towards the final aspirations of Semantic Web, putting people and documents back into the roles as first actors and supporting them with Semantic Web technologies and standards. This is the case for *Semantic Publishing*.

Semantic Publishing concerns "anything that enhances the meaning of a published journal article [more generally, a document], facilitates its automated discovery, enables its linking to semantically related articles, provides access to data within the article in actionable form, or facilitates integration of data between papers" (Shotton 2009). The Semantic Publishing approach goes beyond the current interest of recognising relevant entities in the text and/or transforming natural language statements into formal assertions. In fact, Semantic Publishing aims at describing the

S. Peroni, *Semantic Web Technologies and Legal Scholarly Publishing,*
Law, Governance and Technology Series 15, DOI 10.1007/978-3-319-04777-5_7,

entire discourse and argumentation of (bibliographic) documents through formal tools and semantic technologies. The final aim is to increase the *users' comprehension* of documents through software and applications that work "intelligently" on the formal conceptualisation of the narrative of the documents themselves.

To realise this vision, the actors involved—i.e., publishers, authors, readers, archivists, legal experts, technologists and developers—must be part of an organised cooperative community. Given the intrinsic heterogeneity of the actors involved, Semantic Publishing must be addressed from different perspectives.

Rather than explore the social and economical aspects of Semantic Publishing, I have in this book focussed on its technological environment, presenting examples from the legal scholarly publishing. In order to link a text to the formal representation of its meaning and thus to represent its argumentative discourse, Semantic Publishing needs at least two distinct resources: on the one hand, a powerful and expressive document markup language that allows semantic characterisations of its elements and content. On the other hand, shared models (ontologies) that allow the formal description of all the aspects of a document, from its structure to its argumentative discourse.

My contributions in this direction are shown in two projects: *EARMARK* and *SPAR*. As illustrated in Chaps. 3 and 4 EARMARK is a markup metalanguage that allows one to create markup documents as sets of OWL assertions, without the structural and semantic limits imposed by meta-markup languages such as XML. EARMARK is a platform to link the content layer of a document with its intended formal semantics. Having EARMARK as a solid base for defining the content of documents and its syntactical organisation, I then developed the *Semantic Publishing And Referencing* (*SPAR*) ontologies (Chap. 5), a collection of formal models providing an upper semantic layer for describing the publishing domain. SPAR is a set of eight modular and interoperable ontologies that precisely describe the whole publishing domain using terms from publishers' vocabulary: ranging from bibliographic, structural and rhetorical descriptions of documents to specification of publishing workflows. Thus, using EARMARK as a foundation for SPAR descriptions opens up to a semantic characterisation of all the aspects of a document and of its parts.

Of course, these two aspects—the *markup* and the *semantics*—must be understood, discussed, developed and used within an heterogeneous community that includes people who do not care about the technologies, but who are extremely competent in their own specific domains. Being domain experts, they know the needs and constraints of their own communities. Thus, their contributions to the development of sophisticated Semantic Publishing technologies are crucial.

However, such people may have difficulties in interacting with the Semantic Publishing technologies. Thus, we need user-friendly interfaces that shield such users from the underlying formalisms and semantic models of such technologies.

This is the reason why a good half of my research has concerned the development of interfaces that hide the complexity of markup and ontology formalisms behind user-friendly views. The tools I presented in Chap. 6—*LODE*, *KC-Viz*, *Graffoo* and *Gaffe*—will hopefully find extensively use for presenting ontologies to publishers, for developing new ontologies to meet particular needs, and for allowing authors to add semantic data to their own documents. These tools have had a crucial role in the development of the SPAR ontologies themselves. Without doubt, they have facilitated the frequent and productive interactions I had with publishers and domain experts, and have provided one of the main reasons for the early adoption of SPAR in the publishing domain, as described in Sect. 5.7. LODE and KC-Viz are currently being used in even broader domains and they have been flagged as important contributions in the Semantic Web community[1].

My future research will cover further aspects of Semantic Publishing. In the following sections I introduce the planned future works for all the languages, models and tools presented in this book.

7.1 EARMARK: Future Works

The main and urgent future development of my work on EARMARK, concerns a study of the applicability of this approach to markup in different research domains. In particular, my aim is to investigate real use-case scenarios that involve researchers of different disciplines, such as Humanities or Law. For instance, a relevant issue in Humanities is the use of overlapping markup structures to represent differences among different copies of the same manuscript as a unique digital document. This particular branch of Philology, called *textual criticism*, aims at reconstructing the original text of a manuscript starting from an analysis held on multiple copies of it written by different scribes. Although TEI (Text Encoding Initiative Consortium 2013) enables one to store all the overlapping fragments of a *critical edition* via XML workarounds (introduced in Sect. 2.1.1), my interest is to investigate how different approaches to overlapping markup such as EARMARK can address this problem. Since an interaction with humanists and other researchers that may not be expert in markup technologies is needed, I plan to develop a user interface that facilitates the specification of overlapping markup in EARMARK.

Although I have already carried out a first comparison between XML approaches to overlap and EARMARK (introduced in Sect. 3.2.3), it may be interesting to develop a complexity-based comparison as well, using a richer and more heterogeneous set of input documents. Moreover, this set of documents will be useful for the evaluation of a conversion framework, called the *EARMARK framework*, I am currently developing with my research group. The main aim of the EARMARK framework is to enable the automatic conversion of XML documents with overlapping markup from a format (e.g., ODT) into another (e.g., OpenXML). The framework has been developed so as to use EARMARK as intermediate format to apply the conversion. Part of this work has been already done and introduced in Barabucci et al. (2012).

[1] For instance, LODE is listed in the W3C wiki page about tools for semantic data, available at http://www.w3.org/2001/sw/wiki/LLDtools. Moreover, KC-Viz is now part of the core components of the NeOn Toolkit.

7.2 SPAR: Future Works

Although the SPAR ontologies are already being used within different communities (Sect. 5.7), my prior interest is to empirically evaluate the goodness of all the eight ontological modules to assess the quality of their vocabulary and their ease of use. I also plan to carry out other formal evaluations to understand the quality of those ontologies according to their logical organisation (e.g., through OntoClean (Guarino and Welty 2002) and similar frameworks).

Moreover, I am currently working on the release of triplestores of bibliographic information compliant with SPAR. In particular, in addition to the work already done with the JISC OpenCitation project (Sect. 2.5.1), my research group and I are collaborating with the publishing house *Società Editrice il Mulino*. Our aim is to study a way to enhance its bibliographic objects through SPAR-based semantic assertions and then to publish them as open linked data. Along the lines of the work with the above publishing house, David Shotton (University of Oxford) and I are now managing with Mulberry Technologies Inc.[2] an alignment strategy between their *Journal Article Tag Sets* (*JATS*)[3]—i.e., a set of XML DTDs to store journal articles—and SPAR entities (Peroni et al. 2012).

Another interesting aspect of my proposed research will be the study and development of algorithms for the automatic or semi-automatic identification of structural and rhetoric characteristics of document parts, such as citations (Di Iorio et al. 2013a, b) and other components. Starting from a pattern-based description of a markup document (as introduced in Sect. 5.4.1), it should be possible to deduce the structural roles of its components (sections, chapters, paragraphs, figures, etc., as sketch out in Di Iorio et al. (2013c) as well as their rhetoric functions (introduction, background, experiment, results, etc.) without having an *a priori* knowledge of the intended meaning of such markup elements. My aim is to develop automatic mechanisms that assign structural patterns and DoCO characterisations (Sect. 5.4) to markup elements of XML and EARMARK documents.

7.3 LODE: Future Works

I plan to extend the functionalities of the tool with new features. In particular, future versions of LODE will support full multi-language documentation and the explicit handling of the OWL 2 DL meta-modelling capabilities (i.e., OWL punning) in entity descriptions. Moreover I plan to use the KC-Viz abstraction capabilities, introduced in Sect. 6.3.1, to highlight the most important classes of an ontology in its HTML documentation as rendered through LODE and to develop two plugins, one for the NeOn Toolkit and the other for Protégé, to use LODE within ontology development applications.

[2] Mulberry Technologies Inc.: http://www.mulberrytech.com.

[3] Journal Article Tag Sets: http://www.mulberrytech.com/JATS/.

7.4 KC-Viz: Future Works

From the study of the user questionnaires discussed in Sect. 6.3.3 interesting ideas for future works arose. Although KC-Viz is already integrated within the NeOn Toolkit, some criticisms came up with its integration with other plugins, in particular with those supporting the reasoning and query infrastructure of the NeOn Toolkit. I plan to work on extending KC-Viz in order to enable the key-concept extraction mechanism and navigation according to both the declared and inferred ontological axioms.

Moreover, I plan to increase the interface behaviours of KC-Viz by adding layout mechanisms for coarse-grained views (or sky views) to show the entire ontology, by extending the snapshot feature to handle multiple loaded snapshots at the same time, and by highlighting connected links for a class when it is selected.

7.5 Graffoo: Future Works

There are several planned future developments of Graffoo, but the main priority will be given to its empirical evaluation. In particular, I am interested in understanding whether and how much Graffoo diagrams enable users to understand and develop ontologies. To this end, I plan to carry out a user-testing session that involves people of different fields (Semantic Web practitioners, computer scientists, humanists, etc.) interested in ontologies. The aim is to understand whether Graffoo widget are enough to make sense of a first informal presentation of an ontology.

Besides that, I also plan to work on possible prototypical applications based on the Graffoo widget. First of all, I want to develop a set of XSLT stylesheets to extend DiTTO (Gangemi and Peroni 2013), so as to enable the automatic conversion of a set of *yEd* documents specifying Graffoo diagrams into OWL 2 DL ontologies. The next step will be to develop and implementing a pure Web-based editors for the development and publication of Graffoo diagrams as OWL 2 DL ontologies.

7.6 Gaffe: Future Works

I plan to carry out several evaluation studies to assess the advantages of using Gaffe as authoring tool and form editor according to different kinds of users (ontologists, publishers, semantic data publishers, etc.). In particular, I am now developing a first user-testing session that aims at investigating the benefits introduced by Gaffe when ontologists use it to develop Web forms through the specification of instance documents that link domain ontologies to ontological descriptions of the forms. My aim is to demonstrate how experts in ontology development can make real and usable Web forms despite their inexperience in the development of Web interfaces.

Moreover, I plan to extend the Gaffe API in order to use the systems in different environment such as word processor. In particular I am now designing an integrated system to support users when enriching documents through SPAR. The idea is to use Gaffe as the main interface by which users can associate semantic metadata to markup documents defined through EARMARK, keeping track of provenance information (e.g., through the W3C Provenance Ontology (Sahoo and McGuinness 2013) of both the author(s) of the formal semantic statements about the text and the author(s) of the text itself.

References

Barabucci, G., S. Peroni, F. Poggi, and F. Vitali. 2012. Embedding semantic annotations within texts: The FRETTA approach. In Proceedings of the 27th symposium on applied computing (SAC 2012), 658–663. New York: ACM. doi:10.1145/2245276.2245403.

Berners-Lee, T., J. Hendler, and O. Lassila. 2001. The semantic web. In Scientific American, 17 May 2001. http://www.scientificamerican.com/article.cfm?id=the-semantic-web. Accessed 30 July 2013.

Di Iorio, A., A. Nuzzolese, and S. Peroni. 2013a. Characterising citations in scholarly documents: The CiTalO framework. In *ESWC 2013 satellite events—Revised selected papers*, eds. P. Cimiano, M. Fernàndez, V. Lopez, S. Schlobach, and J. Völker, lecture notes in computer science, vol. 7955, 66–77. Berlin: Springer. doi:10.1007/978-3-642-41242-4_6.

Di Iorio, A., A. Nuzzolese, and S. Peroni. 2013b. Towards the automatic identification of the nature of citations. In Proceedings of 3rd workshop on semantic publishing (SePublica 2013), CEUR workshop proceedings, vol. 994, eds. A. García Castro, C. Lange, P. Lord, and R. Stevens, 63–74. Aachen: CEUR-WS.org. http://ceur-ws.org/Vol-994/paper-06.pdf. Accessed 30 July 2013.

Di Iorio, A., S. Peroni, F. Poggi, D. Shotton, and F. Vitali. 2013c. Recognising document components in XML-based academic articles. Proceedings of the 2013 ACM symposium on document engineering (DocEng 2013), 181–184. New York: ACM. doi:10.1145/2494266.2494319.

Gangemi, A., and S. Peroni. 2013. DiTTO: Diagrams transformation into OWL. In Proceedings of the ISWC 2013 posters & demonstrations track, CEUR workshop proceedings, vol. 1035, eds. E. Blomqvist and T. Groza, 5–8. Aachen: CEUR-WS.org. http://ceur-ws.org/Vol-1035/iswc2013_demo_2.pdf. Accessed 30 July 2013.

Guarino, N., and C. Welty. 2002. Evaluating ontological decisions with OntoClean. *Communications of the ACM* 45 (2): 61–65. doi:10.1145/503124.503150.

Peroni, S., D. A. Lapeyre, and D. Shotton. 2012. Mapping JATS to RDF using the SPAR (semantic publishing and referencing) ontologies. Proceedings of the Journal Article Tag Suite Conference 2012 (JATS-Con 2012). Bethesda: National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/books/NBK100491/. Accessed 30 July 2013.

Sahoo, S., and D. McGuinness. 2013. The PROV ontology. W3C recommendation 30 April 2013. World Wide Web Consortium. http://ceur-ws.org/Vol-1035/iswc2013_demo_2.pdf. Accessed 30 July 2013.

Shotton, D. 2009. Semantic publishing: The coming revolution in scientific journal publishing. *Learned Publishing* 22 (2): 85–94. doi:10.1087/2009202.

Text Encoding Initiative Consortium. 2013. TEI P5: Guidelines for electronic text encoding and interchange. Charlottesville: TEI Consortium. http://www.tei-c.org/Guidelines/P5. Accessed 30 July 2013.