

Chapter 2

The Digital Publishing Revolution

Abstract In this Chapter I discuss the theories and technologies that take part in today's publishing revolution, a.k.a. semantic publishing. In particular, I introduce some of the most important research works on my primary fields of interest, namely markup models and languages to enhance published documents (e.g., legislative documents) semantically, and ontologies/metadata schema to describe such documents. Finally, after introducing some significant research areas in the semantic publishing domain, I conclude the chapter by listing events (i.e., projects, workshops, journal issues, competitions) that have characterised the initial development of the discipline of semantic publishing.

In this Chapter I will discuss the most relevant research areas in semantic publishing. I will focus particularly on my primary fields of interest: *markup* models, which enable the addition of *semantics* within published documents; and document *metadata schemas* and *ontologies*.

First of all, in order to appreciate the general context in which my work is set, it may be useful to describe briefly the changes that took place in digital publishing during the last decade. Scholarly authoring and publishing are undergoing a revolution due to the potential for development coming, on the one hand, from the use of Web-related technologies (e.g., transport protocols, markup languages, the Semantic Web) as a medium of communication, and, on the other hand, from the adoption of new publishing and editorial processes which seem to be converging to a fully-open accessibility of editorial contents and metadata.

The first step of this revolution was made possible by the creation of the Web, which made publishers recognise the digitalisation process and, consequently, the online publication as new effective ways of bibliographic publication. As it had been predicted (Odlyzko 2002), the social and research impact of the availability of scholarly material online is continuing to grow. One of the main reasons of this growth has been the introduction of the Open Access (OA) publishing paradigm¹. Through it, publishers can either directly—the *gold* OA—or indirectly—the *green* OA (Harnad et al. 2004)—publish articles online and offer their complete and free-of-charge worldwide readability and accessibility at no cost.

¹ Probably, the first formal document that used the words “open access” was (Bromley 1991).

Originally, the use of OA was considered a gamble with small chance of success. However, earlier works, such as Lawrence (2001), Harnad and Brody (2004), Swan (2009), gave empirical evidence of the advantages of OA in terms of better visibility, findability and accessibility for research articles. These factors and the development of clear and established strategies (Solomon 2008; Bjork and Hedlund 2009) to change the publishers' business model from a non-OA service to an OA publishing process are some of the most important reasons of the success of the OA model—and of the (increasing) growth and consensus in digital online publication. Moreover, innovative publishing approaches have been recently proposed—e.g., the Liquid Publications Project (Simon et al. 2011) which show how to use Web technologies and OA principles to change and improve the current publication process.

Obviously, we have only covered only the first part of the long way towards a successful and widely accepted Web-oriented digitalisation and publishing of bibliographic materials. In fact, publishers have not adopted Web standards for their work yet. Rather, they still employ a variety of proprietary XML-based informational models and document type definitions (DTDs) (Beck 2010). While such independence was understandable in the pre-Web world of paper publishing, that now appears anachronistic, since publications from different sources and their metadata are incompatible, requiring hand-crafted mappings to convert from one to another. For a large community such as publishers, the lack of standard definitions that could be adopted and reused across the entire industry represents losses in terms of money, time and effort.

In contrast, modern web information management techniques employ standards such as RDF (Carroll and Klyne 2004) and OWL 2 (Motik et al. 2012) to encode information in ways that permit computers to query metadata and integrate web-based information from multiple resources in an automated manner. Since the processes of scholarly communication are central to the practice of science, it is essential that publishers now adopt such standards to permit inference over the entire corpus of scholarly communication represented in journals, books and conference proceedings. This requires the availability of appropriate ontologies and tools that are specially tailored to the requirements of authors, publishers, readers, librarians and archivists.

Some time ago, some research institutes and companies involved in publishing research started to consider whether and how Web technologies could address the issues described above. In retrospect, that moment can be marked as the beginning of what we call today *semantic publishing*.

Semantic publishing is the use of Web and Semantic Web technologies to enhance a published document such as a journal article, in order to enable the definition of formal representations of its meaning; facilitate its automatic discovery; enable its linking to semantically related articles; provide access to data within the article in actionable form; and allow integration of data between papers (Shotton et al. 2009; Shotton 2009). As confirmed by a number of recent initiatives², semantic publishing

² The various initiatives that have been involved research communities around the topics and issues of Semantic Publishing, e.g., the Elsevier Grand Challenge and the SePublica Workshops, will be introduced in Sect. 2.5.

and scholarly citation using Web standards are currently two of the most interesting topics within the scientific publishing domain. We identify some significant research areas in this domain, which include:

1. the *development of markup technologies* that facilitate the creation of complex and semantically-enhanced markup documents, which make possible to have, simultaneously, a formal semantic description of their structures (e.g., chapters, introduction, paragraphs) as well as of their content;
2. the *development of semantic models* (vocabularies, ontologies) that meet the requirements of scholarly authoring and publishing;
3. the *development of visualisation and documentation tools* that permit such ontologies to be easily understood by users who are neither experts nor technicians of particular modelling languages;
4. the *development of annotation tools* that allow these models to be used by end-users (e.g., publishers, editors, authors) for enhancing documents with relevant semantic assertions;
5. the *development of new algorithms* that can take advantages of this new semantic layer of annotations, for example when searching over large sets of on-line documents;
6. the *development of new business models* that arrange effective publishing processes for the creation, use and dissemination of semantic assertions;
7. the *study and realisation of empirical evaluations* that ascertain the benefits and/or the drawbacks of Semantic Publishing for both authors and publishers, such as understanding whether its use increases the impact factor of articles and/or the amount of visits on publishers' Web pages;
8. the *organisation of events*, such as conferences, workshops, projects, journal issues, in order to publicise and promote semantic publishing principles and advantages to a broader audience.

In the rest of this Chap. I will outline related works where solutions to the issues indicated in the first two points of the above list are presented, and that represent the research areas in which my work is set. I will address point three and four in Chap. 6. I will conclude this chapter by listing a series of events (i.e., projects, workshops, journal issues, competitions) that have characterised the initial development of the discipline of semantic publishing.

2.1 Towards Semantics-Aware Markup Languages

The original definition of markup clearly states that markup is used for saying *something* about the content of a document (Coombs et al. 1987). Understanding what “something” refers to is strictly dependant on the particular *semantics* adopted by the markup vocabulary thereby considered. However, markup languages such as XML and SGML do not provide any mechanism to define or associate a particular semantics to their markup structures. Often, the semantical characterisation of such markup

lies between the human's subjective interpretation of the name given to markup items and their natural language descriptions written by the author of a particular markup schema. On the other hand, both the previous examples do not provide a formal and mathematical characterisation (e.g., by means of logic formulas) of such semantics.

Anyway, *overlapping* markup structures are needed when different agents associate multiple (even discording) semantics to the same document fragment. Note that having two different interpretations of a particular document passage is possible, in particular within a domain—i.e., Semantic Publishing—where the analysis and formalisation of the scientific discourse are encouraged. Thus, the topic of overlapping markup, that has been discussed and investigated for years, becomes extremely significant in this context.

In the following sections I will introduce previous studies about both overlapping markup and markup semantics, which are two of the most interesting topics in markup research.

2.1.1 *Overlapping Markup*

The need for multiple overlapping structures over documents using markup syntaxes such as XML and SGML is a long-standing issue, and a large amount of literature exists about techniques, languages and tools that allow users to create multiple entangled hierarchies over the same content. A good review can be found in DeRose (2004).

Some of this research proposes to use plain hierarchical markup (i.e., XML) and employ specially tailored elements or attributes to express the semantics of overlapping in an implicit way. The TEI Guidelines (Text Encoding Initiative Consortium 2013) present a number of different techniques that use SGML/XML constructs to force multiple hierarchies into a single one, including:

- *milestones*, through which one hierarchy is expressed using the standard hierarchical XML markup and the elements belonging to the other ones are represented through a pair of empty elements denoting the start and the end tags, and connected to each other by special attributes;
- *flat milestones*, that represents each of the hierarchy elements as a milestone, i.e., an empty element placed where the start or the end tag should be, all of them contained as children of the same root element;
- *fragmentation*, in which one hierarchy (the primary) is expressed through the standard hierarchical XML markup, while the elements of the secondary hierarchies are fragmented within the primary elements, in a way that suits the primary hierarchy and are connected to each other by special attributes;
- *twin documents*, in which each hierarchy is represented by a different document that contains the same textual content while marking up the elements according to the individual hierarchy;

- *stand-off markup*, which places all the textual content in a single structure with the possible specification of the shared hierarchy, while putting the remaining elements in other structures (e.g., files) with the positional association of each starting and ending location to the main structure, realised by using, for instance, XPointer (DeRose et al. 2001) locations³.

Given the large number of techniques to deal with overlapping structures in XML, in Marinelli et al. (2008), Marinelli et al. present a number of algorithms to convert XML documents with overlapping structures from and to the most common approaches, as well as a prototype implementation.

In Riggs (2002), Riggs introduces a slightly different technique for fragmentation within XML structures. In this proposal, *floating elements*, i.e., those elements that do not fall in a proper or meaningful hierarchical order, are created using the name of the element followed by an index referring to its semantically-related parent element. For example, the floating element `<name.person[2]>John</name.person[2]>` means that `<name>John</name>` is semantically child of the second occurrence of the element *person*, even though the floating element is not structurally contained by its logical parent.

Other research even proposes to abandon of the theory of trees at the base of XML/SGML altogether, and use different underlying models and newly invented XML-like languages that allow the expression of overlaps through some kind of syntactical flourishing.

For instance, *GODDAG* (Sperberg-McQueen and Huitfeldt 2004) is a family of graph-theoretical data structures to handle overlapping markup. A *GODDAG* is a Direct Acyclic Graph whose nodes represent markup elements and text. Arcs are used to explicitly represent containment and father-child relations. Since multiple arcs can be directed to the same node, overlapping structures can be easily represented in *GODDAG*. Full *GODDAGs* cannot be linearised in any form using embedded markup, but *restricted GODDAGs*, a subset thereof, can and have been linearised into *TexMecs* (Huitfeldt and Sperberg-McQueen 2003; Marcoux 2008), a multi-hierarchical markup language that also allows full *GODDAGs* through appropriate workarounds, such as virtual elements.

LMNL (Tennison and Piez 2002) is a general data model based on the idea of *layered text fragments and ranges*, where multiple types of overlap can be modelled using concepts drawn from the mathematical theory of intervals. Multiple serialisations of *LMNL* are available, such as *CLIX* and *LMNL-syntax*.

XConcur (Schonefeld and Witt 2006) is a similar solution based on the representation of multiple hierarchies within the same document through *layers*. Strictly related

³ Note that the use of standoff approaches to handle overlapping issues is not only a prerogative of the world of Computer Science and document markup: it has been in fact adopted also in several projects in Linguistics related to the processing and annotation of natural language texts, for instance the *General Architecture for Text Engineering (GATE)* (Cunningham 2002) and *Callisto* (Day et al. 2004).

to its predecessor CONCUR as it was included in the SGML, XConcur was developed in conjunction with the validation language XConcur-CL to handle relationships and constraints between multiple hierarchies.

The *variant graph* approach (Schmidt and Colomb 2009) is also based on graph theory. Developed to deal with textual variations—that generate multiple versions of the same document with multiple overlapping hierarchies—this theory proposes a new data model to represent literary documents and a graph linearisation (based on lists) that scales well even with a large number of versions. Schmidt et al. recently presented an extension of their theory that also allows users to merge multiple variants into one document (Schmidt 2009).

In Portier and Calabretto (2009) a detailed survey about overlapping approaches was presented, together with a discussion on the MultiX 2 data model—that uses W3C standard languages such as XInclude to link and fetch text fragments within overlapping structures—and a prototype editor for the creation of multi-structured documents.

In Tummarello et al. (2005) a proposal for using RDF as a standoff notation for overlapping structures of XML documents was proposed. By means of the open-source API *RDF Textual Encoding Framework (RDFTef)*, Tummarello et al. demonstrate a possible way for handling overlapping markup within documents and identifying textual content of a document as a set of independent RDF resources that can be linked mutually and with other parent resources.

In addition to giving the opportunity to define multiple structural markup hierarchies over the same text content, the use of RDF as the language for encoding markup allows the user to specify semantic data on textual content as well. However, the main advantage of RDF is the possibility of using particular built-in resources that describe different kinds of containers, either ordered (`rdf:Seq`) or unordered (`rdf:Bag`), as defined in the RDF syntax specification (Carroll and Klyne 2004). Thus, RDF resources can be used to represent every printable element in the text—words, punctuation, characters, typographical symbols, and so on—while RDF containers can be used to combine such fragments and containers as well.

RDF does not provide any mechanism to define a formal vocabulary for structural markup, since it is able neither to define certain resources as classes of a particular kind (elements, attributes, comments, text nodes) nor to characterise the possible relations that such resources may have among others. However the specification of an RDFS (Brickley and Guha 2004) or of an OWL (Motik et al. 2012) layer can successfully address this issue. Hybrid solutions obtained by mixing different models, even when they are built one upon another, may seem elegant but are not necessarily the best choice. In fact, there exist well-known interoperability limits between OWL 2 DL and RDF (Krotzsch et al. 2011) that prevent the correct use of Semantic Web tools and technologies. In particular:

- any markup document made using RDF containers (e.g., to describe what the markup items contain and in which order) and OWL ontologies (e.g., to define classes of markup entities and their semantics) results in a set of axioms that make the OWL ontology completely inconsistent. This limits the applicability

of the most frequently used Semantic Web tools, that are usually built upon the (computationally-tractable) description logic underlying OWL 2 DL;

- the individual analysis of each language may be not applicable when we have to check particular properties that lay between RDF and OWL layers have to be checked. For example, to verify the validity of a markup document against a particular schema, which is one of the most common activities with markup, one needs to work with both markup item structures (that would be defined in RDF) and logical constraints about classes of markup items (e.g., elements only, attributes only, the element “p”, all the element of a particular namespace, etc., all of them definable in OWL).

Being able to express everything we need directly in OWL addresses both issues in a straightforward way. The well-known absence of containers and sequences in OWL can be overcome by modelling classes in specific ways using specific design patterns such as Ciccarese et al. (2008) and Drummond et al. (2006).

2.1.2 Markup Semantics and Semantic Markup

The advent of the Semantic Web (and social web) has induced a shift of meaning for some terms that are traditionally associated with markup languages. Originally, the act of *marking up* was strictly associated with document markup, where the term “tag” was used to refer to *markup elements*: syntactic items representing the building blocks of a document structure. While, in the original definition, markup “tells us something about [the text or content of a *document*]” (Coombs et al. 1987), in the Semantic Web the term “markup” is sometimes used to identify any data added to a *resource* with the intention to semantically describe it (as well as “metadata” or “resource description”). Because of this recent re-drawing of the markup meaning, the term “tag” has also drastically changed its definition to “a non-hierarchical keyword or term assigned to a piece of information (such as an Internet bookmark, digital image, or computer file)”⁴.

Partially because of this shift in meaning—that brought, as first consequence, the fact of having two different (and often unrelated) visions of the Web: the *Web of documents* and the *Web of data*—the Semantic Web has not considered in detail the issue of *markup semantics* (e.g., what is the meaning of a markup element *title* contained in a document *d*?), concentrating all its efforts in dealing with *semantic markup* (e.g., the resource *r* has the string “Semantic enhancement of document markup” as title) (Renear et al. 2002).

However, markup semantics is a very well-known and relevant issue for markup languages and consequently for digital libraries. Nowadays, a large amount of content stored in digital libraries is encoded with XML. XML, as any markup (meta)language,

⁴ http://en.wikipedia.org/wiki/Tag_%28metadata%29

provides a machine-readable mechanism for defining document structure, by associating labels to fragments of text and/or other markup. This association has a particular meaning, since each markup element asserts something about its content. However, what is asserted by the markup is not an issue of the markup itself. One of the goals of markup metalanguages is to avoid imposing any particular semantics: they express mere syntactic labels on the text, leaving the implicit semantics of the markup to the interpretation of humans or tools programmed by humans. Of course, a lot of markup languages, such as HTML, TEI and DocBook, are accompanied by natural language descriptions of their markup, but those descriptions are not machine-readable; in other words, there is no formal mechanism to embed markup semantics within markup language schemas.

Previous works (Renear et al. 2002; Renear et al. 2003; Sperberg-McQueen et al. 2009) pointed out some clear advantages in having a mechanism to define a machine-readable semantics of markup languages: enabling parsers to perform both syntactic and *semantic validation* of document markup; *inferring facts* from documents automatically by means of inference systems and reasoners; simplifying the *federation*, *conversion* and *translation* of documents marked up with different and non-interoperable markup vocabularies; allowing users to *query* upon the structure of the document considering its semantics; creating *visualisations* of documents by considering the semantics of their structure rather than the specific vocabulary in which they are marked up; increasing the accessibility of documents' content, even in the case of *tag abuse* (Dubin 2003), i.e., "using markup languages construction in ways other than" the ones "intended by the language designer"; promoting a more flexible *software design* for those applications that use markup languages, guaranteeing a better *maintainability* even when markup language schemas evolve.

For instance, it could be interesting to query documents for specific XML structures (e.g., all data tables in a collection of scientific papers written by a specific author, regardless of the fact that they were marked up with different vocabularies), or verifying semantic constraints of XML elements regardless of their position within the document (e.g., the utterer of each instance of the speech fragments as transcribed in a parliamentary debate document is uniquely assigned to the individual that purportedly made the speech).

Although the Semantic Web could directly address XML semantics in order to gather the above-mentioned advantages, the Semantic Web community has always considered XML only as a serialisation language for RDF or OWL, or as a way to encode relational data to be subsequently extracted and expressed in RDF. However, these two usages depart from the original goal of XML, i.e., to provide a mechanism for marking up digital documents (books, papers, messages, etc.). Consequently, for example, it is often the case that relational data in XML encode both domain and document semantics; in such cases, extracting semantics from markup by means of bulk recipes generates semantic issues, because the dataset and/or ontologies obtained from that extraction will be unreliable (due to the usually conflicting data/text implicit semantics). A case study of this heterogeneity is the translation of FAO

FIGIS document management schemata⁵, which generates an ontology describing real world entities as well as documents, provenance, interfaces, versioning data, etc.

There is a large literature concerning semantics applied to markup. One of the first attempts for describing formal markup semantics is introduced in Sperberg-McQueen et al. (2000). The basic idea of Sperberg-McQueen et al. is to point out how users apply markup: through it, they make inferences about the document structures and the text those structures contain. According to Sperberg-McQueen et al., “the meaning of markup is the set of inferences it licenses”. The general framework they developed to associate semantics to markup and to make inferences on it needs some representation of the markup document, a *sentence skeleton* for each item of the markup language under consideration in order to associate a meaning, and a set of (categorised) predicates and rules for allowing inferences. In this work, all the examples are illustrated using Prolog both for the representation of the nodes and for defining/inferring semantics using predicates and rules.

Focusing on the best-known meta-markup language, XML, in Renear et al. (2003), Renear et al. discuss problems characterising schema languages for XML, from DTD to XMLSchema: those languages only permit a clear definition of the language syntax, and some of them (RelaxNG (Clark 2001), XML Schema (Gao et al. 2012)) allow the declaration of a simple semantics on the datatypes, and little more. Although annotations can be specified for XMLSchema structures, there is no pre-defined semantics associated to them. Everything else concerning semantics—the meaning of an element, the relationships among items, etc.—is not expressible in a machine-readable format through those schema languages. The Renear et al. propose the BECHAMEL Project as a possible solution to express markup semantics. As they explain in Renear et al. (2002), BECHAMEL allows one to associate semantics with markup by adding new hierarchies to the original structure of the document. Using these additional hierarchies, one can define the meaning of the elements and properties that cannot be expressed using the schema languages alone.

A different approach is used in Simons et al. (2004). Simons et al. developed a framework to associate semantics with any XML document D in a three-step process:

1. defining an OWL ontology O to express all the meanings they want to use;
2. writing a set of rules R in a specific XML language to associate those meanings to a set of elements D ;
3. through a XSLT transformation, processing D using O and R , so obtaining a new semantically-enriched XML document.

Similarly, other works, such as Nuzzolese et al. (2010), Garcia and Celma (2005), Van Deursen et al. (2008), propose a general process that, starting from an XMLSchema S , an XML document D (written according to S) and an ontology O (that can be generated starting from S), allows one to convert all the data in D , described by XML elements and attributes, into appropriate RDF instances consistent with O .

⁵ <http://www.fao.org/fi/figis/devcon/diXionary/index.html>

The approach introduced in Marcoux (2006), Marcoux and Rizkallah (2009) does not provide a formal machine-readable specification for defining markup semantics, but it is useful when human interpretation is needed in structuring a document. Marcoux et al. describe *Intertextual Semantics*, a mechanism to associate meaning with markup elements and attributes of a schema as natural language constructs; this is realised by associating a pre-text and a post-text with each of them. When the vocabulary of a schema is used correctly, the markup content is combined with the pre-text and post-text descriptions to make a correct natural language text that describes the entire information contained in a document. The difference between the common natural language documentation and Intertextual Semantics is that in the latter the meaning of a markup item is dynamically added when writing a document, and, as a consequence, can be read sequentially in the document editor itself.

Of course, eRDF⁶ and RDFa (Adida et al. 2013) may be valid choices for associating—and extracting by means of GRDDL (Connolly 2007) applications—formal semantics with arbitrary text fragments, and to markup elements within documents. Although they are very helpful for annotating documents and adding semantic information about markup elements and their content, their use is possible only by adding new attributes or, even worse, new elements, therefore changing the document structure. The problem being that the need of modifying the document structure is not easily suitable for domains, for example within organisations that deal with administrative or juridical documents, which must always preserve their original structure.

2.2 Markup Languages for Legal and Legislative Documents

Markup languages to describe and define legislative documents have been an hot topic within Computer Science and Law communities for years, and they continue to attract the interests of international entities, such as countries' governments and standardisation institutes. In this section I will introduce only some of the most interesting and famous works in this area.

2.2.1 *Formex*

The *Formalized Exchange of Electronic Publications (Formex)*⁷ (Guittet 1985) is a markup format released by the *Publications Office of the European Union*⁸ in 1985. It was developed to enable data exchanging between the Publication Office and its contractors.

⁶ eRDF: <http://www.ezeneva.ch/w3c-RDF-ResourceDescriptionFramework>.

⁷ Formex homepage: <http://formex.publications.europa.eu>.

⁸ Publication Office of the European Union homepage: <http://publications.europa.eu>.

Originally developed as an SGML-based markup language, the current fourth version (dated May 2004) is totally based on XML technologies, as described in the introductory page of its documentation⁹:

XML is only the starting point for the development of depending standards which allow to transform and, in particular, to present the instances (XSLT, XSLFO). Another important effort in this context was made by developing a new standard for the specification of grammars. DTDs are replaced by XML Schemas which also offers the possibility to define the contents of elements and attributes.

The current specification, called Formex 4 and written entirely in English, uses, thus, XML as base metamarkup language, XML Schema to specify the formal grammar of the language and Unicode UTF-8 to encode Formex documents. In particular the specification consists of two specific parts:

- the physical specifications which contain information on the exchange of data, the construction of filenames and, in particular, on the set of characters set;
- the grammar for the markup based on XML Schema.

2.2.2 *Norme in Rete*

The Italian project *Norme in Rete* (*NIR*, in English *Norms on the Net*) (Marchetti et al. 2002) started in 1999, was led by the Italian Ministry of Justice and financed by the *Autorità per l'Informatica nella Pubblica Amministrazione* (AIPA, in English *Italian Authority for the Information Technology*)¹⁰. The goal of this project was twofolds. Firstly, its partners aimed at creating a freely-accessible Web portal containing all the legal and legislative documents produced by the Italian Parliament, so as to have a centralised access point to all the documents having legal validity in Italy. Second, they wanted to develop an XML-based markup language to store all the legislative documents in a format that facilitated the development of search and annotation tools to easily access and browsing the huge set of normative documents produced by Italian institutions.

NIR includes two different kinds of schemas (developed in DTD and XMLSchema), one *strict* (used for the final version of norms) and the other *loose* (to define drafting rules expressed in form of *circular*). Even though they define the same set of elements, what really changes is the content model of such elements of the particular version in consideration. In particular, the loose schema relaxes several constraints introduced in the strict schema. This makes the documents written according to the strict schema still valid against the loose schema, while the vice versa does not hold.

⁹ Excerpt from <http://formex.publications.europa.eu/formex-4/physspec/formex-4-introduction-.htm>.

¹⁰ Note that AIPA was transformed in *Centro Nazionale per l'Informatica nella Pubblica Amministrazione* (CNIPA) in 2003, and then in *Ente nazionale per la Digitalizzazione della Pubblica Amministrazione* (DigitPA) in 2009. Its current official website is <http://www.digitpa.gov.it/>.

Both DTDs allow the user to define three different kinds of documents:

- articolato con preambolo, i.e., a document that includes a formalised hierarchical structure preceded by a preamble;
- articolato senza preambolo, i.e., a document like the above, but without the preamble;
- semi-articolato, i.e., a document that does not have any particular formal hierarchical structure.

In addition, both DTDs can be split in three different sub-schemas, each defining, respectively, the structure of norms, the organisation of text and all the meta-information associated to the document.

2.2.3 *LexDania*

LexDania XML (Petersen 2005; Lupo et al. 2007) is a joint project held by the Danish Parliament and the Danish Ministry of Justice. The aims of this project are:

- the development of an XML format to write and store legislative documents (mostly acts and rules);
- the development of tools (e.g., editors) to support the creation and editing of such legislative documents in an easy and intuitive way;
- the proposal for a process that allows a gradual transition from a digital storage of legislative documents stored as Word files to a digital storage containing such documents in the aforementioned XML format.

The LexDania XML format is based on and is currently part of the OIOXML-standard for exchanging text documents between agencies in the Danish public administration. The overall design structure of this format is split in three different and interconnected layers, each defining and increasing level of specificity of its markup vocabulary. The first level, called *meta-schema*, defines the general syntax definitions of the language and its core data types. The second level, called *omni-schema*, extends the previous schema enriching it with domain-specific terms and definitions. Finally, the schemas in the third layer, called *application schemas*, extends again the previous schema defining specific document type semantics according to the particular context they want to describe.

LexDania is already in use in several legislative processes of the Denmark Parliament, as documented in official Parliament websites¹¹ and Petersen (2011).

¹¹ E.g., <http://www.ministerialtidende.dk/Forms/L0500.aspx?page=5>.

2.2.4 METALex NL

One of the outputs of the projects E-POWER (IST Project 2000-28125) and e-COURT (IST Project 2000-28199) was a proposal for a Dutch Legal XML Standard (Boer et al. 2002), which was later named METALex¹² (Boer et al. 2002).

METALex is described by its authors as a “generic and easily extensible framework for the XML encoding of the structure and contents of legal and paralegal documents” (Boer et al. 2002). It is composed by several XML Schema documents which define its formal structures. The core of the language is simple. In fact, in order to achieve a sort of independence of jurisdiction, its authors chose to define only those elements that all the regulatory documents of different jurisdiction share. Of course, it is still up to the final user either to use METALex as it is or to appropriately extend it to meet specific requirements derived from specific domain or usage.

In addition to its extendibility, it is also possible to include the textual content of the document in different native languages, using the XML attribute *xml:lang*, so as to have multi-lingual versions of the same document stored within the same file. It is also possible to adapt the vocabulary of the markup language according to the particular target language simply by defining a language-specific schema extension to the neutral vocabulary of the standard document schema. The idea is that the extension should import the standard document schema and should substitute element names with the appropriate names in the target language by specifying it through the attribute *substitutionGroup* of the schema.

2.2.5 CEN MetaLex

Born as evolution of the METALex format (Boer et al. 2002) presented in Sect. 2.2.4, *CEN MetaLex*¹³ (Boer et al. 2007; Boer et al. 2008) is an interchange format defined as XML-based markup language, which aims to be the minimum baseline for other standards for legal and legislative documents. In particular, it was developed not in order to replace such standards with one more format, but rather to propose standardised way of describing legal documents and facilitate their exchange and interoperability.

One of the most important innovations in CEN MetaLex compared with its oldest versions described in Sect. 2.2.4, is the adoption of the *FRBR* specification (International Federation of Library Associations and Institutions Study Group on the Functional Requirements for Bibliographic Records 2009), which I will present in more details in Sect. 2.3.5, to describe legal documents from different abstract and physical perspectives. In particular CEN MetaLex is concerned primarily with the identification of legal bibliographic entities on the basis of their content (i.e., the *Expression* level in FRBR), while it imposes an XML-based language as a mandatory format for storing documents.

¹² METALex homepage: <http://www.metalex.nl>.

¹³ CEN MetaLex homepage: <http://www.metalex.eu>.

CEN MetaLex is defined through a single XML Schema file. Each XML element is characterised by a name, providing a clear meaning for the text fragment it contains, a set of attributes containing additional information about the content of the element itself, and a particular content model according to seven main types: *container* (containing a sequence of other elements and no text), *hcontainer* (as the previous, but with specific elements identifying titles and numbers), *mcontainer* (containing a sequence of other mcontainers and metas, and no text), *block* (containing a sequence of other elements and text nodes), *inline* (as the previous one, but it can be child of blocks or other inline only), *milestone* (containing no text and no elements, and it can be contained by blocks and inlines only). In addition, it is possible to specify metadata associated to the documents using the element *meta* (containing no text and no elements, and it can be contained by mcontainers only) in combination with RDFa (Adida et al. 2013).

2.2.6 Akoma Ntoso

Originally thought to be the standard markup language for e-Parliament services in a Pan-African context and currently primary topic of the OASIS LegalDocumentML (LegalDocML) TC¹⁴, the *Architecture for Knowledge-Oriented Management of African Normative Texts using Open Standards and Ontologies*, a.k.a. *Akoma Ntoso*¹⁵ (Barabucci et al. 2009, 2010), is an XML vocabulary for legal and legislative documents whose primary objective is to allow one to enrich a legal text with semantic data.

Akoma Ntoso focuses on the identification of three main aspects of legal documents:

- the structures composing the document, to be marked up according a precise XML vocabulary based on common structural patterns found in legal documents;
- the references to other related legal documents, made by using a common naming convention based on URIs;
- the storage of non-authoritative annotations, by means of other ontologically-like approaches compatible with Topic Maps (SC34/WG 2003), OWL and GRDDL (Connolly 2007).

The XML documents created according to the Akoma Ntoso specifications use a layered structure where each layer addresses a single problem. First, the *text layer* provides a faithful representation of the original content of the legal text. Then, the *structure layer* provides a hierarchical organisation of the parts present in the text

¹⁴ OASIS LegalDocumentML (LegalDocML) TC homepage: http://www.oasisopen.org/committees/tc_home.php?wg_abbrev=legaldocml

¹⁵ Akoma Ntoso homepage: <http://www.akomantoso.org/>.

layers. Finally, the *metadata layer* associates information from the underlying layers with ontological information. In addition, whenever this semantic information is the result of a subjective interpretation, Akoma Ntoso allows multiple and independent opinions to be stored in a formal way within the same document, and used alternatively, cumulatively or compared to each other.

I will describe Akoma Ntoso in Sect. 4.1 in more details.

2.2.7 *HTML + RDFa and XML in gov.uk Websites*

On the basis of the experience of the US government about increasing public access to datasets generated by the Executive Branch of the Federal Government¹⁶, the UK Government's project *data.gov.uk*¹⁷ (Sheridan and Tennison 2010) was launched in January 2010 with the aim of making datasets, containing data coming from several UK Government departments (9426, up-to-date as of May 19, 2013), freely-available. All data are non-personal (for privacy reasons) and, in principle, available in several formats such as TXT CVS, XLS, RDF and RDFa.

Each page which describes a dataset presents information about the formats used, its openness, its themes, and the temporal coverage of the dataset itself. Some of the informations is also described in embedding RDF statements within HTML pages through RDFa¹⁸ (Adida et al. 2013). The final goal of this project is twofolds. On the one hand, it aims to make governmental data freely-accessible online to the public. On the other hand, it seeks to integrate such data within the Open Linked Data¹⁹, which already counts several datasets coming from different UK companies and institutions (Shadbolt et al. 2012).

One of those institutions, within the legislative domain, is the London Gazette²⁰, which is the most important official journal of record of the British Government. The London Gazette publishes all its material as PDF files and HTML+RDFa pages²¹. The main part of the semantic assertions described through RDFa conforms to the *Gazette Ontology*²², which defines all the classes and properties used for all the Gazette Notices.

¹⁶ Data.gov homepage: <http://www.data.gov>.

¹⁷ Data.gov.uk homepage: <http://data.gov.uk>.

¹⁸ Data Catalog Vocabulary/RDFa in data.gov.uk:
http://www.w3.org/egov/wiki/Data_Catalog_Vocabulary/RDFa_in_data.gov.uk.

¹⁹ An interesting description of lessons learned by the process of conforming Open Government Data with the Open Linked Data is given in Shadbolt et al. (2012).

²⁰ London Gazette homepage: <http://www.london-gazette.co.uk>.

²¹ An example of an HTML+RDFa page in the London Gazette is available at <http://www.london-gazette.co.uk/issues/58664/notices/497223/date=2008-04-10>.

²² Gazette Ontology: <http://www.gazettes-online.co.uk/ontology#>.

A sister project of the previous one, i.e., *legislation.gov.uk*²³, managed by The National Archives²⁴ on behalf of Government of the United Kingdom²⁵, has recently released all UK Legislation from 1267 to the present in several formats²⁶: HTML, XML, RDF and Atom. In particular, an XML Schema was developed²⁷ so as to enable users to retrieve an XML representation of legislative documents according to the *Legislation Schema*²⁸, which permits the specification of both metadata by means of Dublin Core (that will be introduced in Sect. 2.3.1), and the content of legislation using XHTML (W3C HTML Working Group 2002) for tables and MathML (Carlisle et al. 2010) for formulae.

2.3 Metadata Schema, Vocabularies and Ontologies for Publishing

The definition of vocabularies and ontologies that enable the description of document metadata is crucial for the Semantic Publishing. A large number of these metadata schemas appeared in the nineties, and only in recent times their Semantic Web versions were developed either as RDF/RDFS vocabularies or OWL ontologies.

Several vocabularies and/or models for the publishing domain have been developed in the past few years. In this section I will specifically list those that are usually adopted and currently defined through Semantic Web languages and technologies.

2.3.1 Dublin Core

Developed as a result of a conference held in Dublin, Ohio, USA in 1995 that involved both technicians (librarians, publishers, archivists) and academics (researches, software developers), the current versions of Dublin Core (DC) Metadata Elements (Dublin Core Metadata Initiative 2012b) and of DC Metadata Terms (Dublin Core Metadata Initiative 2012a) are the most widely used vocabularies for describing and cataloguing resources.

²³ Legislation.gov.uk homepage: <http://www.legislation.gov.uk>.

²⁴ The National Archives homepage: <http://www.nationalarchives.gov.uk>.

²⁵ Government of the United Kingdom homepage: <https://www.gov.uk>.

²⁶ Formats: <http://www.legislation.gov.uk/developer/formats>.

²⁷ XML format: <http://www.legislation.gov.uk/developer/formats/xml>.

²⁸ Legislation Schema: <http://www.legislation.gov.uk/schema/legislation.xsd>.

These vocabularies have become particularly important and relevant for sharing metadata about documents among different repositories (Koutsomitropoulos et al. 2008) and digital libraries (Montoya et al. 2005), as well as being used to describe documents in HTML (Dublin Core Metadata Initiative 2008; DocBook Walsh 2010) and other XML formats such as Open Document (OpenOffice document format) (JTC1/SC34 WG 6 2006).

While very useful for the creation of basic metadata for resource discovery, the main limitation of DC is a direct consequence of the **generic nature of its terms**. For example, using DC Terms one can identify a creator but not an author; a bibliographic resource but not a journal article; an identifier but not an ISSN, and a date but not a publication date.

2.3.2 PRISM

The *Publishing Requirements for Industry Standard Metadata (PRISM)* (International Digital Enterprise Alliance 2009) is a specification defining a rich set of metadata terms for describing published work. It was developed to address the need of publishers to address emerging requirements for metadata sharing and aggregation, and currently it involves some of the most important publishers and associated companies, such as *Adobe Systems*, the *McGraw-Hill Companies*, *Reader's Digest*, *Time Inc.*, the *Nature Publishing Group*, and *U.S. News and World Report*.

The PRISM metadata terms are expressible both in XML, according to a specific DTD, and in RDF (Hammond 2008). These terms are explicitly recommended for the specification of metadata of documents expressed through markup languages such as (DocBook Walsh 2010). Moreover, these terms are also included in ontologies describing the publishing domain, such as the *Bibliographic Ontology (BIBO)*²⁹ (D'Arcus and Giasson 2009), which is discussed below.

While PRISM has a much richer set of terms that describe bibliographic entities than DC, its main limitation is that it is a **flat structure**, lacking hierarchies. That prevents its use for a complete description of the characteristics of bibliographic entities. For example, while the data property *prism:volume* permits the volume number of a bibliographic reference to be represented as a string, PRISM lacks the concept of "Volume" as a distinct class among other bibliographic classes that have a hierarchical partitive relationship to one another (i.e., Journal Article > Issue > Volume > Journal), and whose members can possess other properties, such as having authors and editors.

²⁹ BIBO, the Bibliographic Ontology: <http://purl.org/ontology/bibo/>.

2.3.3 BIBO

BIBO, i.e., the *Bibliographic Ontology* (D’Arcus and Giasson 2009), is an OWL Full ontology that allows one to write descriptions of documents (*bibo:Document* is the core class of that model) for publication on the Semantic Web. It includes both DC terms (Dublin Core Metadata Initiative 2012a) and PRISM (International Digital Enterprise Alliance 2009) to cover common needs, and it adds other classes and properties to describe in more detail the publishing domain, such as *bibo:AcademicArticle*, *bibo:Journal*, *bibo:Collection*, *bibo:Book*, *bibo:Chapter* and *bibo:Issue*. BIBO is a good ontology that is widely used among the bibliographic community.

From a pure computational perspective, BIBO defines the range of the property *bibo:authorList* using either an *rdf:List* or an *rdf:Seq*, therefore making the model non-compliant with OWL 2 DL. That limits the applicability of reasoners and other Semantic Web tools that are usually built upon the (computationally-tractable) description logic underlying OWL 2 DL.

2.3.4 MARC 21

Another relevant work in this field, widely used in the Libraries community and developed before the introduction of the Semantic Web, is the *MARC 21 Format for Bibliographic Data* (Library of Congress - Network Development and MARK Standard Office 2010). Introduced in 1961, MARC 21 is a very complex code for describing bibliographic resources as one of seven different primary types: book, continuing resource, computer file, maps, music, visual materials and mixed materials. To each resource, there can be associated different kinds of metadata, such as titles, names, subjects, notes, publication data, etc.

In MARC21, each type of metadata is represented by a three-digit code (called a *tag* in the MARC21 specification) that identifies the main metadata category of relevance. Other characters can follow this tag in order to specify additional information. For example, let me introduce a simple bibliographic reference describing Casanovas et al. (2007):

Pompeu Casanovas, Núria Casellas, Christoph Tempich, Denny Vrandečić, Richard Benjamins (2007). OPJK and DILIGENT: ontology modeling in a distributed environment. <http://link.springer.com/content/pdf/10.1007%2Fs10506-007-9036-2.pdf>.

To express these data in MARC21 I will need to use the following tags:

```

100 1#$aCasanovas , Pompeu
100 1#$aCasellas,Núria
100 1#$aTempich , Christoph
100 1#$aVrandečić, Denny
100 1#$aBenjamins , Richard
260 ##$c2007
145 10$aOPJK and DILIGENT: ontology modeling in a
    distributed environment
856 40$uhttp://link.springer.com/content/pdf/10.1007%2
    Fs10506-007-9036-2.pdf

```

where “100” identifies a personal name, “260” indicates the year of publication, “145” the title of a work, and “856” the electronic location of that entity.

With the advent of the Semantic Web, MARC21 was formalised as an RDF vocabulary (Styles et al. 2008) in order to be adopted and used in Semantic Web applications. However, many librarians now regard MARC as too complex and esoteric, and are undergoing a mind shift to more pragmatic open standards.

2.3.5 FRBR

The *Functional Requirements for Bibliographic Record (FRBR)* (International Federation of Library Associations and Institutions Study Group on the Functional Requirements for Bibliographic Records 2009) is a general model, proposed by the International Federation of Library Association (IFLA), for describing documents and their evolution. It works for both physical and digital resources and it has proved to be very flexible and powerful. One of the most important aspects of FRBR is the fact that it is not associated with a particular metadata schema or implementation.

The following brief description outlines FRBR’s basic concepts and the way they can be applied within a publishing domain. FRBR describes all documents from four different and correlated points of view: *Work*, *Expression*, *Manifestation* and *Item*; each of which is a FRBR *Endeavour*. These can be illustrated by considering of the book *Alice’s Adventures in Wonderland* by Lewis Carroll as an example:

- **Work.** A FRBR *Work* is a high-level abstract Platonic concept of the *essence* of a distinct intellectual or artistic creation, for example the ideas in Lewis Carroll’s head concerning *Alice’s Adventures in Wonderland*, independent of any representation of these ideas in a particular form. A Work is realised through one or more Expressions;
- **Expression.** A FRBR *Expression* is the realisation of the intellectual or artistic *content* of a Work. Thus the original text of *Alice’s Adventures in Wonderland* and its Italian translation *Le Avventure di Alice nel Paese delle Meraviglie* refer to different Expressions of the same Work. An Expression is embodied in one or more Manifestations;

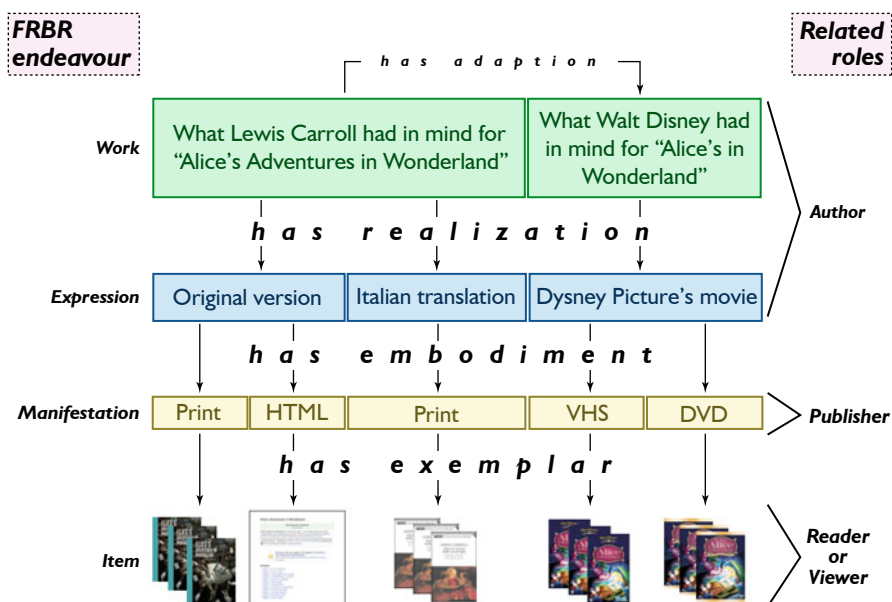


Fig. 2.1 The four FRBR layers, with a specification of roles that people may play in each layer

- **Manifestation.** A FRBR *Manifestation* of a work defines its particular physical or electronic embodiment, for example, the particular *format* in which "Alice's Adventures in Wonderland" is stored: as a printed object or in HTML, represent two quite different Manifestations. In publishing, different manifestations of a journal article will all bear the same Digital Object Identifier (DOI), which identifies the Expression of the work, not its various Manifestations. However, a paperback and a hardback version of a book will bear different International Standard Book Numbers (ISBNs), since these identifiers are assigned at the Manifestation level. A Manifestation is exemplified by one or more Items;
- **Item.** A FRBR Item is a particular physical or electronic *copy* of *Alice's Adventures in Wonderland* that a person can own, for example the printed version of the book you have in your bookcase, or the Mobipocket format copy you have downloaded to read on your e-book device. All Items that are identical to one another—for example books from the same printing, are exemplars of the same Manifestation.

In Fig. 2.1, I summarise the four distinct FRBR layers with particular reference to the publishing domain, using as example *Alice's Adventures in Wonderland*, and I indicate the most common roles (*Author*, *Publisher* and *Reader/Viewer*) that usually people have with respect to each layer.

Despite the increased expressivity enabled by these layers, the greatest limitation of FRBR with respect to the publishing domain is its lack of terms **that permit publications to be described in normal everyday language** (e.g., "research paper",

“review”, “book chapter”, “newspaper editorial”) rather than using the more abstract and esoteric FRBR-specific terms “work”, “expression”, “manifestation” and “item”.

A further limitation that FRBR has in common with other standards—i.e., DC (Dublin Core Metadata Initiative 2012b; Dublin Core Metadata Initiative 2012a), PRISM (International Digital Enterprise Alliance 2009)—is that it has hitherto been implemented and shared **only as XML or RDF vocabularies**, rather than as OWL DL ontologies, preventing it from being used within applications that employ reasoning based on description logic models.

There now exist two different implementations of the core concepts of FRBR using standards that permit the encoding of proper formal semantics: the first, authored in 2005 by Richard Newman and Ian Davis in RDFS³⁰ and the second, developed from the first, created in 2010 by Paolo Ciccarese and I in OWL 2 DL³¹ (Ciccarese and Peroni 2013).

In addition to being particularly adequate for the description of scholarly works, FRBR can be used also in the context of the legislative process of many different legal systems. For instance, in Civil Law legal systems, the modification in time of a legal text that has been approved in the past is a common practice, and thus it is crucial to be able to keep track of the way such text is modified. In this case, we could use an FRBR-based approach that represents the legal document as a whole through a FRBR Work, while each consecutive re-writing due to a modification of the current approved version could be specified by an FRBR Expressions explicitly linked to the other ones through a temporal relation (e.g., *is revision of* or *is successor of*).

Similarly, FRBR can be very useful within Common Law legal systems, such as the federal legislation of the United States, where it is crucial to keep track of the process of law codification, where, for instance, a certain statute is restated as positive law in the US Code. Although this topic is partially out of the scope for this book (as I mainly deal with legal scholarly publishing), I will introduce an example of such scenario in Sect. 5.6.2.

2.3.6 RDA

The *Resource Description and Access (RDA)*³² is a standard released in June 2010 by the American Library Association³³, the Canadian Library Association³⁴, and the Chartered Institute of Library and Information Professionals (CILIP)³⁵ in the UK.

³⁰ The FRBR Core in RDFS: <http://vocab.org/frbr/core>.

³¹ The FRBR Core in OWL 2 DL: <http://purl.org/spar/frbr>.

³² RDA, the Resource Description and Access, standard homepage: <http://www.rda-jsc.org/rda.html>.

³³ American Library Association homepage: <http://www.ala.org/>.

³⁴ Canadian Library Association homepage: <http://www.cla.ca/>.

³⁵ Chartered Institute of Library and Information Professionals homepage: <http://www.cilip.org.uk/>.

RDA allows one to describe resources related to libraries, archives, museums, and other organisations working on information management of bibliographic entities and cultural heritage artefacts.

RDA is the the official replacement of the *Anglo-American Cataloguing Rules Second Edition (AACR2)*³⁶, and it is based on the Functional Requirements for Bibliographic Records (FRBR), introduced in Sect. 2.3.5. Essentially, RDA is a format-independent approach that standardises how metadata should be identified and structured.

RDA takes particular attention to possible ways of interaction with Semantic Web applications and the Linked Open Data. To this end, during the development of the standard, a parallel task force has been investigating how to develop formal and machine-readable representations of RDA element sets and value vocabularies so as to be used by both humans and machines. This activity resulted in the release of a set of RDF-based vocabularies (Hillmann et al. 2010) freely available online³⁷. Each element set and vocabulary is available as a descriptive HTML page as well as in RDF/XML format (Beckett 2004). The element sets define the main FRBR concepts as formal OWL classes and support the relations existing between them, as well as an extended list of agent roles organised according to the FRBR layers. Additional RDA vocabularies define value sets for different applicative contexts (e.g., values for digital formats, picture colouring, text extents, etc.).

2.3.7 SWAN Citations Ontology

Another model previously used to define bibliographic resources is the *Citations Ontology*³⁸ included in the SWAN (Semantic Web Applications in Neuroscience) Ontologies, version 1.2 (Ciccarese et al. 2008). In this ontology, the main class *Citation*³⁹ is used as super-class, of which all the other resources (e.g., *JournalArticle*, *WebArticle* and *Book*) are sub-classes.

The main advantage of this ontology is that it is completely developed in OWL 2 DL. Contrary to BIBO, which defines the range of the property *bibo:authorList* using either an *rdf:List* or an *rdf:Seq* and therefore makes the model non-compliant with OWL 2 DL, the Citation Ontology uses the SWAN Collections Ontology module⁴⁰. This is an OWL 2 DL ontology that allows one to handle lists of authors and contributors of a bibliographic object, thus enabling the specification of ordered lists while still keeping the ontology locally consistent.

³⁶ Anglo-American Cataloguing Rules Second Edition homepage: <http://www.aacr2.org/>.

³⁷ The RDA (Resource Description and Access) Vocabularies: <http://rdvocab.info/>.

³⁸ SWAN Citations Ontology Module: <http://swan.mindinformatics.org/spec/1.2/citations.html>.

³⁹ Note that in SWAN the concept “Citation” is used to represent the cited object itself, rather than the performative act of making a citation.

⁴⁰ SWAN Collections Ontology Module: <http://swan.mindinformatics.org/spec/1.2/collections.html>.

The main problem of the Citations Ontology is the sparseness of its vocabulary, and the difficulty of aligning it with other structurally complex models such as FRBR, because, as with BIBO, it collapses all bibliographic entity descriptions within the single class *Citation*.

2.3.8 SKOS

Publishers need to classify the documents they publish according to discipline-specific thesauri or classification schemes, for example those belonging to economics⁴¹ or computer science⁴².

The *Simple Knowledge Organization System* (SKOS) (Miles and Bechhofer 2009) is an RDFS model which supports the use of knowledge organisation systems (KOS) such as thesauri, classification schemes, subject heading lists and taxonomies within the framework of the Semantic Web. The reception of this language has been particularly positive: a large number of well-known thesauri and classification systems have started to convert their respective specifications into SKOS documents^{43, 44, 45, 46}. This makes SKOS the *de facto* standard for encoding controlled vocabularies for the Semantic Web.

2.4 Ontologies for Legal Documents

The development of ontologies is one of the most discussed and addressed topic by several communities and players in the legal domain, from philosophers to computer scientists. Several works have been done in the past on this topic, as reviewed by Casellas in Casellas (2011).

Generally speaking, legal ontologies can be clustered according to a bipartite classification:

- core legal ontologies, which provide general definitions of general legal entities such as norm⁴⁷, legal role⁴⁸, legal process⁴⁹;

⁴¹ The Journal of Economic Literature Classification Scheme: http://www.aeaweb.org/jel/jel_class_system.php.

⁴² The Association for Computing Machinery (ACM) Computing Classification System 1998: <http://www.acm.org/about/class/1998>.

⁴³ AGROVOC: <http://aims.fao.org/website/AGROVOC-Thesaurus/sub>.

⁴⁴ The Medical Subject Headings (MeSH): <http://www.ncbi.nlm.nih.gov/mesh>.

⁴⁵ The Library of Congress Subject Headings (LCSH): <http://id.loc.gov/search/>.

⁴⁶ Nuovo Soggettario of the National Central Library in Florence: <http://thes.bncf.firenze.sbn.it/>

⁴⁷ E.g., the class *norm* in <http://www.estrellaproject.org/lkif-core/norm.owl>.

⁴⁸ E.g., the class *legal role* in <http://www.estrellaproject.org/lkif-core/legal-role.owl>.

⁴⁹ E.g., the class *process* in <http://www.estrellaproject.org/lkif-core/process.owl>.

- domain legal ontologies, which define concepts specific to a precise context and/or peculiar to a particular field, such as parliament⁵⁰, judge⁵¹, etc.

Only some of the ontologies in these two sets are developed through Semantic Web technologies (i.e., OWL). For instance, as OWL-based core ontologies, we can cite the *Core Legal Ontology* (Gangemi et al. 2005), the *MetaLex Ontology* (Comité Européen de Normalisation 2010), and *Legal Knowledge Interchange Format core legal ontology* (Breuker et al. 2006). The number of ontologies in the other set (i.e., domain legal ontologies) is undoubtedly larger than the former—we can cite the *BEST ontologies* (van Laarschot et al. 2005), the *Legal Case Ontology* (Wyner and Hoekstra 2012), and the *Parliament Ontology*⁵² as exemplars. We recommend interested readers to see (Casellas 2011) for a more precise discussion of ontologies of both types.

I believe it is appropriate to distinguish, within the first category of ontologies, the subset of legal core ontologies that describe aspects of legal markup languages strictly related to document metadata information—i.e., which have concepts shared among document metadata definitions by the majority of legal and legislative markup languages, such as those introduced in Sect. 2.2—from the others. I call the former subset of ontologies: *ontologies for legal documents*.

Even though the identification of the concepts emerging from the textual content of legal and legislative documents is an important issue to address within the legal domain, the purpose of this work is to discuss how *document metadata* can be represented by and/or linked to some sort of semantic characterisation of them. Thus, even though a discussion of OWL-based legal ontologies may be of interest to the reader, for the aim of this book, I will focus only on ontologies for legal documents.

On these premises, in this section I will briefly introduce some of the ontologies for legal documents (developed in OWL) referenced in Casellas (2011). In Sect. 2.4.4, I will also present an additional ontology, called ALLOT, that was developed specifically to define semantic descriptions of Akoma Ntoso documents (Barabucci et al. 2009). Some of the concepts defined in these ontologies are also used to introduce an example of use of legal markup languages and Semantic Web technologies to assess the quality of legal drafting, as discussed in Sect. 4.4.

2.4.1 *MetaLex Ontology*

CEN MetaLex (introduced in Sect. 2.2.5) defines also an OWL ontology⁵³ describing its core components, i.e., bibliographic elements, document metadata and citations

⁵⁰ E.g., the class *parliament* in <http://reference.data.gov.uk/def/parliament>.

⁵¹ E.g., the class *judge* in <http://www.wyner.info/research/case-ontology.owl>.

⁵² The Parliament Ontology: <http://reference.data.gov.uk/def/parliament>.

⁵³ MetaLex Ontology: <http://justinian.leibnizcenter.org/MetaLex/metalex-cen.owl>.

(Comité Européen de Normalisation 2010). Any *bibliographic object* is actually an instance of one of the classes described in FRBR (International Federation of Library Associations and Institutions Study Group on the Functional Requirements for Bibliographic Records 2009)—i.e., *work*, *expression*, *manifestation* and *item*, as introduced in Sect. 2.3.5—and it is provided with a unique identifier, which is the IRI used to refer to it. In addition, the authors added a particular class, i.e., *bibliographic citation*, to provide a mechanism to refer to a particular bibliographic object from a different context.

On the basis of the MetaLex Ontology, in 2011 Rinke Hoekstra presented an ongoing project to publish all the Dutch legislation published since 2002⁵⁴ as Open Linked Data (Hoekstra 2011). He developed and still maintains the *MetaLex Document Server (MDS)*⁵⁵, a store of 280619338 triples (up-to-date as of May 21, 2013) that describe Dutch national regulations (describing 33702 document versions) in CEN MetaLex XML and as RDF Linked Data. MDS exports its data as RDF triple that conforms to the MetaLex Ontology, which allows the modelling of legislative modification events in terms of time, processes and provenance information regarding the event itself by means of external ontologies such as the Simple Event Model (van Hage et al. 2011), the W3C Time Ontology (Hobbs and Pan 2006) and the Open Provenance Model Vocabulary (Zhao 2010).

Another concrete use of the CEN MetaLex ontology within actual collections of legislative documents is the *legislation.gov.uk* website, collecting all the UK Legislation since 1267. As described in the website⁵⁶, an agent can ask for the RDF/XML representation of legislative works, such as acts, and, thus, can retrieve a CEN MetaLex compliant description of such textual content.

2.4.2 Core Legal Ontology

The *Core Legal Ontology (CLO)*⁵⁷ (Gangemi et al. 2005) is an OWL ontology based on the “Descriptions and Situations” extension to DOLCE (Gangemi and Mika 2003), which is an OWL-based foundational ontology for domain-independent axiomatic theories.

CLO allows one to reason over constraints given by the particular context in consideration, which can be dynamically used by agents “when recognizing or classifying a state of affairs” (Gangemi et al. 2005). In addition, it was developed to address three specific kinds of legal tasks:

⁵⁴ The original data were retrieved from <http://www.wetten.nl>.

⁵⁵ MetaLex Document Server homepage: <http://doc.metalex.eu>.

⁵⁶ RDF/XML format: <http://www.legislation.gov.uk/developer/formats/rdf>.

⁵⁷ Core Legal Ontology: <http://www.loa.istc.cnr.it/ontologies/CLO/CoreLegal.owl>.

- *conformity checking*, which requires to link particular situations (both social and legal ones) to legal norms;
- *legal advice*, which requires to investigate the relationships that exist between legal cases and non-expected situations;
- *norm comparison*, which requires to recognise conflicts between norms according to the particular European and national legislation.

2.4.3 LKIF Core Legal Ontology

The *Legal Knowledge Interchange Format (LKIF) core legal ontology*⁵⁸ (Breuker et al. 2006) is a set of OWL ontologies which aims to describe the legal domain. It includes fifteen different ontology modules, each specialised in a specific legal topic.

Four of these modules—*top*, *place*, *mereology* and *time*—are used to describe the most abstract concepts of the model, such as: top categories, places seen as absolute (e.g., a mountain) and relative (e.g., a ship), description of part-whole relationships, and description of moments and intervals.

Other four models—*process*, *role*, *action* and *expression*—are used to describe basic level concepts, such as processes and changes, types of roles (e.g., epistemic, person, organisation roles), actions performed by particular players, and expressions of attitudes, qualifications and statements.

The aforementioned modules are then extended by three other modules—*legal action*, *legal role* and *norm*—to form the legal core of LKIF. Those modules enable the description of public acts, legal and natural people, legal professions, norms, legal sources and different types of rights and powers.

All modifications and rules are described in two separate modules—*modification* and *rule*—since they represent the two larger frameworks of the legal domain described herein, and thus deserve to be formalised as single modules.

Finally, the first eleven modules described above are integrated within LKIF core module—*lkif-core*—and all of the above modules are included in the LKIF extended module—*lkif-extended*—for a total of fifteen ontology modules.

2.4.4 A Light Legal Ontology on TLC

As briefly introduced in Sect. 2.2.6, Akoma Ntoso (Barabucci et al. 2009, 2010) prescribes the use of ontologies to describe metadata and entities that concern the particular document in consideration. However, Akoma Ntoso does not impose the use of a particular ontological model, and it suggests that ontologies should comply

⁵⁸ LKIF repository: <http://github.com/RinkeHoekstra/lkif-core>.

with fourteen abstract entities called *Top Level Classes (TLC)*: *concept, event, location, object, organisation, person, process, reference, role, term*, and the four FRBR classes (International Federation of Library Associations and Institutions Study Group on the Functional Requirements for Bibliographic Records 2009) *work, expression, manifestation* and *item*.

ALLOT (A Light Legal Ontology on TLC)⁵⁹ (Barabucci et al. 2013) is a lightweight ontology that provides a vocabulary to describe Akoma Ntoso TLC as OWL classes. It can be used to describe in detail the various references that occur in Akoma Ntoso documents, and to map those references to other related entities exposed by means of other models, such as the MetaLex Ontology Sect. 2.4.1 or LKIF Sect. 2.4.3.

In addition, ALLOT is also aligned⁶⁰, when possible, with entities defined in external ontologies, such as DC Metadata Terms (Dublin Core Metadata Initiative 2012a), LKIF Time module⁶¹, SKOS (Miles and Bechhofer 2009), FOAF (Brickley and Miller 2010), PRO (Peroni et al. 2012) and the RDF implementation of the FRBR model (introduced in Sect. 2.3.5).

2.5 Projects, Conferences and Initiatives about Semantic Publishing

Between 2010 and 2013, a large number of initiatives emerged with the precise intention of advertising Semantic Publishing to a broader audience. Each of them, from projects to workshops and journal issues, seemed to confirm that semantic publishing and scholarly citation using Web standards constitute currently two of the most interesting topics within the scientific publishing domain.

In this section I will list the most important initiatives concerning Semantic Publishing in 2010–2013, sponsored by both academia and companies. They represent proof of the increasing interest in Semantic Publishing by scientific, academic and industrial institutions.

2.5.1 JISC's Open Citation and Open Bibliography Projects

In 2010, JISC funded two sister projects: the *Open Citation project*⁶² and the *Open Bibliography project*⁶³, held respectively by the University of Oxford and the University of Cambridge. Their broad goal was to study the feasibility, the advantages

⁵⁹ ALLOT core: <http://akn.web.cs.unibo.it/allot/core>.

⁶⁰ ALLOT implementation: <http://akn.web.cs.unibo.it/allot/impl>.

⁶¹ LKIF time module: <http://www.estrellaproject.org/lkif-core/time.owl>.

⁶² Open Citation project blog: <http://opencitations.wordpress.com>.

⁶³ Open Bibliography project blog: <http://openbiblio.net>.

and the applications at using RDF datasets and OWL ontologies when describing and publishing bibliographic data and citations.

The Open Citation project, in which I was actively involved, intended to increase the effectiveness of scientific publishing and scholarly communication, making available on the Web bibliographic information as RDF data, according to particular ontologies developed for the description of the publishing domain. In particular, the project aimed to create a semantic infrastructure to describe articles as bibliographic records and to report citations between citing articles and cited ones.

The main outcomes of this project are:

- the development of a suite of ontologies, called *Semantic Publishing And Referencing (SPAR)* ontologies, which I developed under the supervision of professor David Shotton when I was based at the University of Oxford. This represents an important part of my work (Chap. 5);
- the development of two tools for ontology documentation, i.e., the *Live OWL Documentation Environment (LODE)* (Peroni et al. 2012), and visualisation, i.e., the *Graphical Framework for OWL Ontologies (Graffoo)*, which I developed to support users when understanding and documenting ontologies. They are introduced in detail in Sect. 6.2 and Sect. 6.4 respectively;
- a corpus of interlinked bibliographic records⁶⁴ obtained converting the whole set of reference lists contained in all the PubMed Central⁶⁵ Open Access articles into RDF data according to SPAR ontologies. The converted RDF data are published as Linked Open Data.

The Open Bibliographic project aimed at publishing a large corpus of bibliographic data as Linked Open Data, starting from four different sources: the Cambridge University Library⁶⁶, the British Library⁶⁷, the International Union of Crystallography⁶⁸, and PubMed⁶⁹. The key strategies promoted by this project were:

- the transformation of publishers' models so as to include the open publication of bibliographic data as Linked Open Data;
- the immediate and continuing engagement of the scholarly community.

⁶⁴ It is available online at <http://opencitations.net>.

⁶⁵ PubMed Central: <http://www.ncbi.nlm.nih.gov/pmc/>.

⁶⁶ Cambridge University Library: <http://www.lib.cam.ac.uk>.

⁶⁷ British Library: <http://www.bl.uk>.

⁶⁸ International Union of Crystallography: <http://www.iucr.org>.

⁶⁹ PubMed: <http://www.ncbi.nlm.nih.gov/pubmed/>.

2.5.2 JISC's Lucero Project

The *Lucero project*⁷⁰ is another JISC project, held by the Open University, which aims to explore the use of Linked Data within the academic domain. In particular, it proposes solutions that could use the Linked Data to connect educational and research content, so that students and researches could benefit from semantic technologies.

Lucero main aims are:

- to promote the publication as Linked Open Data of bibliographic data through a tool to facilitate the creation and use of semantic data;
- to identify a process in order to integrate the Linked Data publication of bibliographic information as part of the University's workflows;
- to demonstrate the benefits derived from exposing and using educational and research data as Open Linked Data, through the development of applications that improve the access to those data.

2.5.3 SePublica and Linked Science

Two of the most important Semantic Web conferences, i.e., the *extended* and the *international* ones, began to promote explicitly Semantic Publishing through two specific workshops.

The workshop *SePublica*⁷¹, co-located within the 8th Extended Semantic Web Conference, is the first formal event entirely dedicated to Semantic Publishing. The aim of the workshop was to investigate upon the different practices of using semantic technologies within the publishing industry. During this half-day workshop were presented seven different papers were presented, one of which was awarded best workshop paper (winning an award of \$750, sponsored by Elsevier). Supported by the success of the first edition, further events followed, *SePublica 2012* and *SePublica 2013*, co-located respectively within the 9th and the 10th Extended Semantic Web Conference.

The *Linked Science*⁷² workshop, a full-day event co-located within the 10th International Semantic Web Conference, involved researchers and practitioners discussing new ways of publishing, sharing, linking and analysing scientific resources, such as articles, datasets and results. Each of the five workshop sessions related to a particular topic, from data-based applications to semantic integration of data. The event concluded with an open meeting in which the various topics of the workshop were

⁷⁰ Lucero project blog: <http://lucero-project.info>.

⁷¹ The 1st International Workshop about Semantic Publication (*SePublica 2011*): <http://sepublica.mywikipaper.org>.

⁷² The 1st International Workshop on Linked Science (*LISC2011*): <http://data.linked-science.org/events/lisc2011>.

discussed. Similarly to SePublica, also the Linked Science workshop was followed by two other editions—Linked Science 2012 and Linked Science 2013, co-located respectively within the 11th and 12th International Semantic Web Conference –, as well as a tutorial⁷³ held at the 18th International Conference on Knowledge Engineering and Knowledge Management.

2.5.4 *Beyond Impact, the PDF and Research Communication*

Recently the interest in proposing and adopting new formats for the improvement of research communications appears obvious. Almost all research and industrial communities agree that the current formats are not sufficient for the needs of Web-based research communication. The workshops entitled *Beyond the PDF*, organised at the University of California San Diego in January 2011⁷⁴ and by FORCE11⁷⁵ in Amsterdam in March 2013, went in that direction. The scope of the event was to identify a set of requirements, applications and deliverables that can be used to accelerate knowledge sharing and discovery.

Beyond the PDF was not the only event organised with the aim of exploring new research possibilities and directions in scholarly publication. The workshop *The Future of Research Communication*⁷⁶, held in Dagstuhl in August 2011, was another gathering where scientists and practitioners coming from different disciplines met to discuss the future direction in scholarly publishing. All the prominent topics of scientific communication were discussed, and, in particular, new formats were proposed, which addressed the changes in media and modes of communication, drew opportune infrastructures and outlined social challenges with the aim of showing possible directions on the future of scholarly communication.

From a broader point of view but always considering the Web as prominent medium of communication, the *Beyond Impact Workshop*⁷⁷ (held in London in May 2011) tried to establish different forms of *impact*—that is a measure of how research outcomes influence and are used by other people—in today's and in future publishing. The output of this workshop is a document⁷⁸ that introduces research collaborations and future works to be done in the next few years.

⁷³ Tutorial on Linked Science 2012 homepage: <http://linkedscience.org/events/tolsci2012>.

⁷⁴ The Beyond the PDF Workshop homepage: <http://sites.google.com/site/beyondthepdf/>.

⁷⁵ FORCE11 homepage: <http://www.force11.org>.

⁷⁶ The Future of Research Communication Dagstuhl Workshop: <http://www.dagstuhl.de/de/programm/kalender/semhp/?seminr=11331>.

⁷⁷ The Beyond Impact Workshop: <http://beyond-impact.org/>

⁷⁸ Beyond Impact Workshop Report: http://docs.google.com/document/d/1sH3JOW5Luki4i37Ve1mOnI2wNZJbaUOx1T42S_7txQ0/edit?hl=en_GB.

2.5.5 *Special Issues of Journals on Semantic Publishing*

Journals about Semantic Web technologies and digital publishing started to be actively interested in Semantic Publishing topics.

An example of that is a special issue of the *Semantic Web Journal*⁷⁹: *New Models of Semantic Publishing in Science*⁸⁰. The central topic addressed by this special issue was about the promotion of new forms of Web-based publications, which allow a rapid and automatic integration of research information, thus making it readily available and reproducible. Towards this goal, the issue identifies Semantic Web technologies as key tools for providing efficient ways to work with new modes of scientific publications and asks for submissions of researches in various related fields: Semantic Publishing, Computer Supported Collaborative Work, Linked Data, eScience and Workflow-driven tools, and Digital Libraries. The editors' aim is to promote and advertise all the important researches that are underway in this field.

Another example of special issue is the *Journal of Web Semantics on Life Science and e-Science*⁸¹. Its goal is to identify which role the semantic technologies play in the context of life sciences such as biology, genetics, zoology, etc. In particular, this special issue want to report on the ways in which semantic technologies can enrich research, publish and reuse data according to semantic-aware formats, increase the scholarly communication, and the like.

References

- Adida, B., M. Birbeck, S. McCarron, and S. Pemberton. 2013. RDFa core 1.1. 2nd ed. Syntax and processing rules for embedding RDF through attributes. W3C recommendation 22 August 2013. World Wide Web Consortium. <http://www.w3.org/TR/rdfa-syntax/>. Accessed 30 July 2013.
- Barabucci, G., L. Cervone, M. Palmirani, S. Peroni, and F. Vitali. 2009. Multi-layer markup and ontological structures in Akoma Ntoso. In *Proceeding of the international workshop on AI approaches to the complexity of legal systems II (AICOL-II)*, lecture notes in computer science 6237 vols, ed. P. Casanovas, U. Pagallo, G. Sartor, and G. Ajani, 133-149. Berlin: Springer. doi:10.1007/978-3-642-16524-5_9.
- Barabucci, G., L. Cervone, A. Di Iorio, M. Palmirani, S. Peroni, and F. Vitali. 2010. Managing semantics in XML vocabularies: An experience in the legal and legislative domain. *Proceedings of balisage: The markup conference 2009*. Rockville: Mulberry Technologies, Inc. <http://www.balisage.net/Proceedings/vol5/html/Barabucci01/BalisageVol5-Barabucci01.html>. Accessed 30 July 2013.
- Barabucci, G., A. Di Iorio, F. Poggi, and F. Vitali. 2013. Integration of legal datasets: From meta-model to implementation. *Proceedings of international conference on information integration and web-based applications and services (iiWAS2013)*, 585-594. New York: ACM. doi:10.1145/2539150.2539180.

⁷⁹ Semantic Web Journal: <http://www.semantic-web-journal.net>.

⁸⁰ New Models of Semantic Publishing in Science: <http://www.semantic-web-journal.net/blog/special-issue-new-models-semantic-publishing-science>.

⁸¹ Special issue of the Journal of Web Semantics on Life Science and e-Science: <http://journalofwebsemantics.blogspot.it/2013/03/cfp-special-issue-on-life-science-and-e.html>.

- Beck, J. 2010. Report from the field: PubMed central, an XML-based archive of life sciences journal articles. Proceedings of the international symposium on XML for the long haul: Issues in the long-term preservation of XML. doi:10.4242/BalisageVol6.Beck01.
- Beckett, D. 2004. RDF/XML syntax specification (Revised). W3C recommendation, 10 February 2004. World Wide Web Consortium. <http://www.w3.org/TR/rdf-syntax-grammar/>. Accessed 30 July 2013.
- Bjork, B., and T. Hedlund. 2009. Two scenarios for how scholarly publishers could change their business model to open access. *Journal of Electronic Publishing* 12 (1). doi:10.3998/3336451.0012.102.
- Boer, A., R. Hoekstra, and R. Winkels. 2002. METALex: Legislation in XML. In Proceedings of the 15th annual conference on legal knowledge and information systems (JURIX 2002), ed. T. Bench-Capon, A. Daskalopulu, and R. Winkels, 1–10. Amsterdam: IOS Press. <http://www.jurix.nl/pdf/j02-01.pdf>. Accessed 30 July 2013.
- Boer, A., R. Hoekstra, R. Winkels, T. Engers van, and F. Willaert. 2002. Proposal for a dutch legal XML standard. In Proceedings of the 1st international conference on Electronic Government (EGOV 2002), lecture notes in computer science. 2456 vols, ed. R. Traummuller and K. Lenk, 142–149. Berlin: Springer. doi:10.1007/3-540-46138-8_22.
- Boer, A., R. Winkels, F. Vitali. 2007. Proposed XML standards for law: MetaLex and LKIF. In Proceedings of the 12th annual conference on legal knowledge and information systems (JURIX 2007), ed. A. R. Lodder and L. Mommers, 19–28. Amsterdam: IOS Press.
- Boer, A., R. Winkels, F. Vitali. 2008. MetaLex XML and the legal knowledge interchange format. Computable models of the law, languages, dialogues, games, ontologies, lecture notes in computer science. 4884 vols, ed. P. Casanovas, G. Sartor, N. Casellas, and R. Rubino, 21–41. Berlin: Springer. doi:10.1007/978-3-540-85569-9_2.
- Breuker, J., A. Boer, R. Hoekstra, and K. van den Berg. 2006. Developing content for LKIF: Ontologies and frameworks for legal reasoning. Proceedings of the 19th annual conference on legal knowledge and information systems (JURIX 2006), ed. T. M. van Engers, 169–174. Amsterdam: IOS Press.
- Brickley, D., and R. V. Guha. 2004. RDF vocabulary description language 1.0: RDF schema. W3C recommendation 10 February 2004. World Wide Web Consortium. <http://www.w3.org/TR/rdf-schema/>. Accessed 30 July 2013.
- Brickley, D., and L. Miller. 2010. FOAF vocabulary specification 0.98. Namespace document, 9 August 2010—Marco polo edition. <http://xmlns.com/foaf/spec/>. Accessed 30 July 2013.
- Bromley, A. 1991. Policy statements on data management for global change research. <http://www.gcric.org/USGCRP/DataPolicy.html>. Accessed 30 July 2013.
- Carlisle, D., P. Ion, and R. Miner. 2010. Mathematical markup language (MathML) version 3.0. W3C recommendation, 21 October 2010. World Wide Web Consortium. <http://www.w3.org/TR/MathML3/>. Accessed 30 July 2013.
- Carroll, J., and G. Klyne. 2004. Resource Description Framework (RDF): Concepts and abstract syntax. W3C recommendation, 10 February 2004. World Wide Web Consortium. <http://www.w3.org/TR/rdf-concepts/>. Accessed 30 July 2013.
- Casanovas, P., N. Casellas, C. Tempich, D. Vrandeic, and R. Benjamins. 2007. OPJK and DILIGENT: Ontology modeling in a distributed environment. *Artificial Intelligence and Law* 15 (2): 171–186. doi:10.1007/s10506-007-9036-2.
- Casellas, N. 2011. Legal ontology engineering. Law, governance and technology series 3. Berlin: Springer. doi:10.1007/978-94-007-1497-7.
- Ciccarese, P., and S. Peroni. 2013. The collections ontology: Creating and handling collections in OWL 2 DL frameworks. To appear in *Semantic Web—Interoperability, Usability, Applicability*. doi:10.3233/SW-130121.
- Ciccarese, P., E. Wu, J. Kinoshita, G. Wong, M. Ocana, A. Ruttenberg, and T. Clark. 2008. The SWAN biomedical discourse ontology. *Journal of Biomedical Informatics* 41 (5): 739–751. doi:10.1016/j.jbi.2008.04.010.

- Clark, J. 2001. RELAX NG specification. Committee specification. Committee specification 3 December 2001. Organization for the advancement of structured information standards. <http://relaxng.org/spec-20011203.html>. Accessed 30 July 2013.
- Comité Européen de Normalisation. 2010. Metalex (Open XML interchange format for legal and legislative resources). CEN workshop agreement 15710:2010 (E). Brussels: Comité Européen de Normalisation. <http://ftp.cen.eu/CEN/Sectors/List/ICT/CWAs/CWA15710-2010-Metalex2.pdf>. Accessed 30 July 2013.
- Connolly, D. 2007. Gleaning resource descriptions from dialects of languages (GRDDL). W3C recommendation 11 September 2007. World Wide Web Consortium. <http://www.w3.org/TR/grddl/>. Accessed 30 July 2013.
- Coombs, J. H., A. H. Renear, and S. J DeRose. 1987. Markup systems and the future of scholarly text processing. *Communications of the ACM* 30 (11): 933–947. doi:10.1145/32206.32209.
- Cunningham, H. 2002. GATE, a general architecture for text engineering. *Computers and the Humanities*, 36(2): 223–254. doi:10.1023/A:1014348124664.
- D’Arcus, B., and F. Giasson. 2009. Bibliographic ontology specification. Specification document, 4 November 2009. <http://bibliontology.com/specification>. Accessed 30 July 2013.
- Day, D. S., C. McHenry, R. Kozierok, and L. Riek. 2004. Callisto: A configurable annotation workbench. In Proceedings of the 4th international conference on Language Resources and Evaluation (LREC 2004), ed. M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, R. Silva, C. Pereira, and S. Barros, 2073–2076. Paris: European Language Resources Association. <http://www.lrec-conf.org/proceedings/lrec2004/pdf/612.pdf>. Accessed 30 July 2013.
- DeRose, S. 2004. Markup overlap: A review and a horse. Proceedings of the extreme markup languages 2004. Rockville: Mulberry Technologies, Inc. <http://conferences.idealiance.org/extreme/html/2004/DeRose01/EML2004DeRose01.html>. Accessed 30 July 2013.
- DeRose, S., E. Maler, and R. Daniel. 2001. XPointer xpointer scheme. W3C working draft, 19 December 2002. World Wide Web Consortium. <http://www.w3.org/TR/xptr-xpointer/>. Accessed 30 July 2013.
- Drummond, N., A. Rector, R. Stevens, G. Moulton, M. Horridge, H. H. Wang, and J. Seidenberg. 2006. Putting OWL in order: Patterns for sequences in OWL. In Proceedings of the workshop on OWL: Experiences and directions (OWLED 2006), CEUR workshop proceedings. 216 vols, ed. B. C. Grau, P. Hitzler, C. Shankey, and E. Wallace. Aachen: CEUR-WS.org. http://ceur-ws.org/Vol-216/submission_12.pdf. Accessed 30 July 2013.
- Dubin, D. 2003. Object mapping for markup semantics. Proceedings of the extreme markup languages 2003. Rockville: Mulberry Technologies, Inc. <http://conferences.idealiance.org/extreme/html/2003/Dubin01/EML2003Dubin01.html>. Accessed 30 July 2013.
- Dublin Core Metadata Initiative. 2008. Expressing dublin core metadata using HTML/XHTML meta and link elements. DCMI recommendation. <http://dublincore.org/documents/dc-html/>. Accessed 30 July 2013.
- Dublin Core Metadata Initiative. 2012a. DCMI metadata terms. DCMI recommendation. <http://dublincore.org/documents/dcmi-terms/>. Accessed 30 July 2013.
- Dublin Core Metadata Initiative. 2012b. Dublin core metadata element set, Version 1.1. DCMI recommendation. <http://dublincore.org/documents/dces/>. Accessed 30 July 2013.
- Gangemi, A., and P. Mika. 2003. Understanding the semantic web through descriptions and situations. In Proceedings of the on the move confederated international conferences, CoopIS, DOA, and ODBASE 2003 (CoopIS/DOA/ODBASE 2003), lecture notes in computer science. 2888 vols, ed. R. Meersman, Z. Tari, and D. C. Schmidt, 689–706. Berlin: Springer. doi:10.1007/978-3-540-39964-3_44.
- Gangemi, A., M. T. Sagri, and D. Tiscornia. 2005. A constructive framework for legal ontologies. Law and the semantic web: Legal ontologies. Methodologies, legal information retrieval, and applications, lecture notes in computer science. 3369 vols, ed. V. R. Benjamins, P. Casanovas, J. Breuker, and A. Gangemi, 97–124. Berlin: Springer. doi:10.1007/978-3-540-32253-5_7.
- Gao, S., C. M. Sperberg-McQueen, and H. S. Thompson. 2012. W3C XML schema definition language (XSD) 1.1 Part 1: Structures. W3C recommendation 5 April 2012. World Wide Web Consortium. <http://www.w3.org/TR/xmlschema11-1/>. Accessed 30 July 2013.

- Garcia, R., O. Celma. 2005. Semantic integration and retrieval of multimedia metadata. Proceedings of the 5th international workshop on knowledge markup and semantic annotation (SemAnnot 2005), CEUR workshop proceedings 185: 69–80. Aachen: CEUR-WS.org. <http://ceur-ws.org/Vol-185/semAnnot05-07.pdf>. Accessed 30 July 2013.
- Guittet, C. 1985. Formex: Formalized exchange of electronic publications. Luxembourg: office for official publications of the European communities. (ISBN: 978-9282553992).
- Hammond, T. 2008. RDF site summary 1.0 modules: PRISM. http://nurture.nature.com/rss/modules/mod_prism.html. Accessed 30 July 2013.
- Harnad, S., and T. Brody. 2004. Comparing the impact of Open Access (OA) vs. Non-OA articles in the same journals. *D-Lib Magazine* 10 (6). doi:10.1045/june2004-harnad.
- Harnad, S., T. Brody, F. Vallieres, L. Carr, S. Hitchcock, Y. Gingras, C. Oppenheim, H. Stamerjohanns, and E. R. Hilf. 2004. The access/impact problem and the green and gold roads to open access. *Serials Review* 30 (4): 310–314. doi:10.1016/j.serrev.2004.09.013.
- Hillmann, D., K. Coyle, J. Phipps, and G. Dunsire. 2010. RDA vocabularies: Process, outcome, use. *D-Lib Magazine* 16 (1/2). doi:10.1045/january2010-hillmann.
- Hobbs, J. R., and F. Pan. 2006. Time ontology in OWL. W3C working draft, 27 September 2006. World Wide Web Consortium. <http://www.w3.org/TR/owl-time/>. Accessed 30 July 2013.
- Hoekstra, R. 2011. The MetaLex document server—legal documents as versioned linked data. In Proceedings of the 10th international semantic web conference (ISWC 2011), Part II, lecture notes in computer science. 7032 vols, ed. L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. F. Noy, and E. Blomqvist, 128–143. Berlin: Springer. doi:10.1007/978-3-642-25093-4_9.
- Huitfeldt, C., and C. M. Sperberg-McQueen. 2003. TexMECS: An experimental markup metalanguage for complex documents. <http://decentius.aksis.uib.no/mlcd/2003/Papers/textmecs.html>. Accessed 30 July 2013.
- International Digital Enterprise Alliance. 2009. Publishing requirements for industry standard metadata specification version 2.0. Alexandria: IDEAlliance. <http://www.idealliance.org/specifications/prism>. Accessed 30 July 2013.
- International Federation of Library Associations and Institutions Study Group on the Functional Requirements for Bibliographic Records. 2009. Functional requirements for bibliographic records final report. International federation of library associations and institutions. http://www.ifla.org/files/cataloguing/frbr/frbr_2008.pdf. Accessed 30 July 2013.
- JTC1/SC34 WG 6. 2006. ISO/IEC 26300:2006—Information technology—open document format for office applications (OpenDocument) v1.0. Geneva: International Organization for Standardization. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=43485. Accessed 30 July 2013.
- Koutsomitropoulos, D. A., G. D. Solomou, and T. S. Papatheodorou. 2008. Semantic interoperability of dublin core metadata in digital repositories. In Proceedings of the 5th international conference on Innovations in Information Technology (IIT 2008), 233–237. Washington: IEEE Computer Society. doi:10.1109/INNOVATIONS.2008.4781709.
- Krotzsch, M., F. Simancik, and I. Horrocks. 2011. A description logic primer. Ithaca: Cornell University Library. <http://arxiv.org/pdf/1201.4089v1>. Accessed 30 July 2013.
- Lawrence, S. 2001. Free online availability substantially increases a paper's impact. *Nature* 411 (6837): 521. doi:10.1038/35079151.
- Library of Congress-Network Development and MARK Standard Office. 2010. MARK 21 format for bibliographic data. 1999 edition, further updates October 2001 and October 2010. <http://www.loc.gov/marc/bibliographic/>. Accessed 30 July 2013.
- Lupo, C., F. Vitali, E. Francesconi, M. Palmirani, R. Winkels, E. de Maat, A. Boer, and P. Mascellani. 2007. General XML format(s) for legal sources. Deliverable 3.1 of the European project for standardised transparent representation in order to extend legal accessibility (ESTRELLA). EU IST-2004-027655. <http://www.estrellaproject.org/doc/D3.1-General-XML-formats-For-Legal-Sources.pdf>. Accessed 30 July 2013.

- Marchetti, A., F. Megale, E. Seta, and F. Vitali. 2002. Using XML as a means to access legislative documents: Italian and foreign experiences. *ACM SIGAPP Applied Computing Review* 10 (1): 54–62. doi:10.1145/568235.568246.
- Marcoux, Y. 2006. A natural-language approach to modeling: Why is some XML so difficult to write? Proceedings of the extreme markup languages 2006. Rockville: Mulberry Technologies, Inc. <http://conferences.idealliance.org/extreme/html/2006/Marcoux01/EML2006Marcoux01.html>. Accessed 30 July 2013.
- Marcoux, Y. 2008. Graph characterization of overlap-only TexMECS and other overlapping markup formalisms. Proceedings of balisage: The markup conference 2008. Rockville: Mulberry Technologies, Inc. <http://www.balisage.net/Proceedings/vol1/html/Marcoux01/BalisageVol1-Marcoux01.html>. Accessed 30 July 2013.
- Marcoux, Y., and E. Rizkallah. 2009. Intertextual semantics: A semantics for information design. *Journal of the American Society for Information Science and Technology* 60 (9): 1895–1906. doi:10.1002/asi.21134.
- Marinelli, P., F. Vitali, and S. Zacchiroli. 2008. Towards the unification of formats for overlapping markup. *New Review of Hypermedia and Multimedia* 14 (1): 57–94. doi:10.1080/13614560802316145.
- Miles, A., and S. Bechhofer. 2009. SKOS simple knowledge organization system reference. W3C recommendation 18 August 2009. World Wide Web Consortium. <http://www.w3.org/TR/skos-reference/>. Accessed 30 July 2013.
- Montoya, E., M. Ruiz, and J. Giraldo. 2005. BDN: A dublin core-based architecture for digital libraries. Proceedings of the international conference on dublin core and metadata applications 2005. Singapore: Dublin Core Metadata Initiative. <http://dcpapers.dublincore.org/pubs/article/view/802/798>. Accessed 30 July 2013.
- Motik, B., P. F. Patel-Schneider, and B. Parsia. 2012. OWL 2 web ontology language: Structural specification and functional-style syntax (Second edition). W3C recommendation 11 December 2012. World Wide Web Consortium. <http://www.w3.org/TR/owl2-syntax/>. Accessed 30 July 2013.
- Nuzzolese, A. G., A. Gangemi, and V. Presutti. 2010. Gathering lexical linked data and knowledge patterns from FrameNet. Proceedings of the 6th international conference on knowledge capture (K-CAP 2011), 41–48. New York: ACM. doi:10.1145/1999676.1999685.
- Odlyzko, A. 2002. The rapid evolution of scholarly communication. *Learned Publishing* 15 (1): 7–19. doi:10.1087/095315102753303634.
- Peroni, S., D. Shotton, and F. Vitali. 2012. Scholarly publishing and the linked data: Describing roles, statuses, temporal and contextual extents. In Proceedings of the 8th international conference on semantic systems (I-SEMANTICS 2012), ed. H. Sack and T. Pellegrini, 9–16. New York: ACM. doi:10.1145/2362499.2362502.
- Peroni, S., D. Shotton, and F. Vitali. 2012. The live OWL documentation environment: A tool for the automatic generation of ontology documentation. In Proceedings of the 18th international conference on knowledge engineering and knowledge management (EKAW 2012), lecture notes in computer science. 7603 vols, ed. A. ten Teije, J. Völker, S. Handschuh, H. Stuckenschmidt, M. d'Aquin, A. Nikolov, N. Aussenac-Gilles, and N. Hernandez, 398–412. Berlin: Springer. doi:10.1007/978-3-642-33876-2_35.
- Petersen, K. E. 2005. Lex dania XML status April 2005. Proceedings of the third workshop on legislative XML: 13–19. Rome: Centro Nazionale per l'Informatica nella Pubblica Amministrazione. http://www.digitpa.gov.it/sites/default/files/Quaderno8XML_0.pdf. Accessed 30 July 2013.
- Petersen, K. E. 2011. Experiences with “Lex Dania Live”. From information to knowledge, frontiers in artificial intelligence and applications. 236 vols, ed. M. A. Biasiotti and S. Faro, 69–76. Amsterdam: IOS Press. doi:10.3233/978-1-60750-988-2-69.
- Portier, P., and S. Calabretto. 2009. Methodology for the construction of multi-structured documents. Proceedings of balisage: The markup conference 2009. Rockville: Mulberry Technologies, Inc. <http://balisage.net/Proceedings/vol3/html/Portier01/BalisageVol3-Portier01.html>. Accessed 30 July 2013.

- Renear, A., D. Dubin, and C. M. Sperberg-McQueen. 2002. Towards a semantics for XML markup. Proceedings of the 2002 ACM symposium on Document Engineering (DocEng 2002), 119–126. New York: ACM. doi:10.1145/585058.585081.
- Renear, A., D. Dubin, C. M. Sperberg-McQueen, and C. Huitfeldt. 2003. XML semantics and digital libraries. Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2003), 303–305. Washington: IEEE Computer Society. doi:10.1109/JCDL.2003.1204879.
- Riggs, K.R. 2002. XML and free text. *Journal of the American Society for Information Science and Technology* 53 (6): 526–528. doi:10.1002/asi.10063.
- SC34/WG3. 2003. Topic Maps. ISO 13250. Geneva: International organization for standardization. <http://www.isotopicmaps.org/>. Accessed 30 July 2013.
- Schmidt, D. 2009. Merging multi-version texts: A generic solution to the overlap problem. Proceedings of balisage: The markup conference 2009. Rockville: Mulberry Technologies, Inc. <http://balisage.net/Proceedings/vol3/html/Schmidt01/BalisageVol3-Schmidt01.html>. Accessed 30 July 2013.
- Schmidt, D., and R. Colomb. 2009. A data structure for representing multi-version texts online. *International Journal of Human-Computer Studies* 67 (6): 497–514. doi:10.1016/j.ijhcs.2009.02.001.
- Schonefeld, O., and A. Witt. 2006. Towards validation of concurrent markup. Proceedings of the extreme markup languages 2006. Rockville: Mulberry Technologies, Inc. <http://conferences.idealliance.org/extreme/html/2006/Schonefeld01/EML2006Schonefeld01.html>. Accessed 30 July 2013.
- Shadbolt, N., K. O'Hara, T. Berners-Lee, N. Gibbins, H. Glaser, W. Hall, and M. C. Schraefel. 2012. Linked open government data: Lessons from data.gov.uk. *IEEE Intelligent Systems* 27 (3): 16–24. doi:10.1109/MIS.2012.23.
- Sheridan, J., and J. Tennison. 2010. Linking UK government data. In Proceedings of the Linked Data on the Web workshop (LDOW 2010), CEUR workshop proceedings. 628 vols, ed. C. Bizer, T. Heath, T. Berners-Lee, and M. Hausenblas. Aachen: CEUR-WS.org. http://ceur-ws.org/Vol-628/ldow2010_paper14.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.
- Shotton, D., K. Portwin, G. Klyne, and A. Miles. 2009. Adventures in semantic publishing: Exemplar semantic enhancements of a research article. *PLoS Computational Biology* 5 (4): e1000361. doi:10.1371/journal.pcbi.1000361.
- Simon, J., A. Birukou, F. Casati, R. Casati, and M. Marchese. 2011. Liquid publications green paper. http://peerevaluation.org/data/ca75910166da03ff9d4655a0338e6b09/PE_doc_28223.pdf. Accessed 30 July 2013.
- Simons, G. F., W. D. Lewis, S. O. Farrar, D. T. Langendoen, B. Fitzsimons, and H. Gonzalez. 2004. The semantics of markup: Mapping legacy markup schemas to a common semantics. In Proceedings of the workshop on NLP and XML (NLPXML 2004), ed. N. Ide and L. Romary, 25–32. Stroudsburg: Association for Computational Linguistics. <http://acl.ldc.upenn.edu/acl2004/nlpxml/pdf/simons-etal.pdf>. Accessed 30 July 2013.
- Solomon, J. S. 2008. Developing open access journals: A practical guide. Oxford: Chandos Publishing Limited. (ISBN: 1843343394).
- Sperberg-McQueen, C. M., and C. Huitfeldt. 2004. GODDAG: A data structure for overlapping hierarchies. In Proceeding of the 5th international workshop on the Principles of Digital Document Processing (PODDP 2000), lecture notes in computer science 2023, ed. P. R. King and E. V. Munson, 139–160. Berlin: Springer. doi:10.1007/978-3-540-39916-2_12.
- Sperberg-McQueen, C. M., C. Huitfeldt, and A. Renear. 2000. Meaning and interpretation of markup. *Markup Languages: Theory and Practice* 2 (3): 215–234. doi:10.1162/-109966200750363599.
- Sperberg-McQueen, C. M., Y. Marcoux, and C. Huitfeldt. 2009. Two representations of the semantics of TEI lite. Proceedings of Digital Humanities 2010 (DH 2010). <http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-663.html>. Accessed 30 July 2013.

- Styles, R., D. Ayers, and N. Shabir. 2008. Semantic marc, MARC21 and the semantic Web. In Proceedings of the WWW 2008 workshop on Linked Data on the Web (LDOW2008), CEUR workshop proceedings. 369 vols, ed. C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee. Aachen: CEUR-WS.org. <http://www.ceur-ws.org/Vol-369/paper02.pdf>. Accessed 30 July 2013.
- Swan, A. 2009. The open access citation advantages: Studies and results to date. Technical report, School of Electronics & Computer Science, University of Southampton. <http://eprints.ecs.soton.ac.uk/18516/>. Accessed 30 July 2013.
- Tennison, J., and W. Piez. 2002. The Layered Markup and Annotation Language (LMNL). Presented at the extreme markup languages conference 2002. 4–9 August 2002, Montreal.
- 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.
- Tummarello, G., C. Morbidoni, and E. Pierazzo. 2005. Toward textual encoding based on RDF. In Proceedings of the 9th ICCCI international conference on Electronic Publishing (ELPUB2005), ed. M. Dobrev and J. Engelen. Leuven: Peeters Publishing Leuven. http://elpub.architecturez.net/system/files/pdf/206elpub2005.content_0.pdf. Accessed 30 July 2013.
- Van Deursen, D., C. Poppe, G. Martens, E. Mannens, R. Van de Walle. 2008. XML to RDF conversion: A generic approach. In Proceedings of the 4th international conference on automated solutions for cross media content and multi-channel distribution (AXMEDIS 08), ed. P. Nesi, J. Delgado, and K. Ng, 138–144. Washington: IEEE Computer Society. doi:10.1109/AXMEDIS.2008.17.
- van Hage W. R., V. Malaisé, R. Segers, L. Hollink, and G. Schreiber. 2011. Design and use of the simple event model (SEM). *Journal of Web Semantics: Science, Services and Agents on the World Wide Web* 9 (2): 128–136. doi:10.1016/j.websem.2011.03.003.
- van Laarschot R., W. van Steenberg, H. Stuckenschmidt, A. R. Lodder, F. van Harmelen. 2005. The legal concepts and the Layman's terms—bridging the gap through ontology-based reasoning about liability. In Proceedings of the 18th annual conference on legal knowledge and information systems (JURIX 2005), ed. M. Moens Katholieke, P. Spyns, 115–125. Amsterdam: IOS Press. <http://www.jurix.nl/pdf/j05-20.pdf>. Accessed 30 July 2013.
- W3C HTML Working Group. 2002. XHTML™ 1.0 The extensible hypertext markup language (Second edition). W3C recommendation, 1 August 2002. World Wide Web Consortium. <http://www.w3.org/TR/xhtml1>. Accessed 30 July 2013.
- Walsh, N. 2010. DocBook 5: The definitive guide. Sebastopol: O'Really Media. (Version 1.0.3. ISBN: 0596805029).
- Wyner, A., and R. Hoekstra. 2012. A legal case OWL ontology with an instantiation of Popov v. Hayashi. *Artificial Intelligence and Law* 20 (1): 83–107. doi:10.1007/s10506-012-9119-6.
- Zhao, J. 2010. Open provenance model vocabulary specification. Revision 1. 0, 6 October 2010. <http://purl.org/net/opmv/ns>. Accessed 30 July 2013.

Semantic Web Technologies and Legal Scholarly
Publishing

Peroni, S.

2014, XXIII, 304 p. 211 illus., 46 illus. in color.,

Hardcover

ISBN: 978-3-319-04776-8