LOD2 for Media and Publishing

Christian Dirschl[1(✉)], Tassilo Pellegrini[2], Helmut Nagy[2], Katja Eck[1], Bert Van Nuffelen[3], and Ivan Ermilov[4]

[1] Wolters Kluwer Deutschland GmbH, Unterschleißheim, Germany
cdirschl@wolterskluwer.de
[2] Semantic Web Company GmbH, Vienna, Austria
[3] TenForce, Leuven, Belgium
[4] Institute of Computer Science, Leipzig University, Leipzig, Germany

Abstract. It is the core business of the information industry, including traditional publishers and media agencies, to deal with content, data and information. Therefore, the development and adaptation of Linked Data and Linked Open Data technologies to this industry is a perfect fit. As a concrete example, the processing of legal information at Wolters Kluwer as a global legal publisher through the whole data life cycle is introduced. Further requirements, especially in the field of governance, maintenance and licensing of data are developed in detail. The partial implementation of this technology in the operational systems of Wolters Kluwer shows the relevance and usefulness of this technology.

Keywords: Data transformation · Data enrichment · Metadata management · Linked data visualization · Linked data licensing · IPR · Wolters Kluwer · Media · Publishing · Legal domain

1 Introduction

1.1 Rationale for the Media and Publishing Use Case

The media and publishing use case within the LOD2 project[1] aims at enabling large-scale interoperability of (legal) domain knowledge based on Linked Data. This is a necessary precondition in the media industry to profit from the benefits of distributed and heterogeneous information sources (DBpedia, EuroVoc) on the Semantic Web. Hence, this use case aims at improving access to high-quality, machine-readable datasets generated by publishing houses for their customers.

This attempt is accompanied by several challenges: Traditional official content such as laws and regulations or court case proceedings are increasingly publicly available on the web and are directly published by the respective issuing bodies. Social networks and platforms, such as Wikipedia, aggregate professional knowledge and publish it at no charge. At the same time e.g. news media generate large amounts of relevant information about events and people that are complementary to conventional content of specialized publishers, but hardly integrated (exception is e.g. integration between BBC and DBpedia[2]). In addition, the amount of relevant information is still growing

[1] http://lod2.eu/Welcome.html, accessed May 10, 2014.
[2] Kobilarov et al. [7].

S. Auer et al. (Eds.): Linked Open Data, LNCS 8661, pp. 133–154, 2014.
DOI: 10.1007/978-3-319-09846-3_7

exponentially; this amount cannot be incorporated and structured by using traditional manual annotation mechanisms. Finally, the customer expects more and more exact and to-the-point information in her actual professional workflow that covers individual interests, personal preferences and one central trusted access to distributed data sources. Interests and preferences of a professional can even change over time and tasks to be completed.

From the perspective of Wolters Kluwer, the relevance of using schema-free data models like RDF and SKOS as well as accessing external content for their data-driven business is obvious.[3] By interlinking quality-approved proprietary data sources and "tapping" classification resources from the community and existing references in the LOD cloud, Wolters Kluwer is exploring diversification scenarios for existing assets as well as business opportunities under new licensing regimes. These efforts must lead to a win-win situation, where, on the one hand, additional revenues can be created by adding value to existing products and, on the other hand, customers of Wolters Kluwer and the public can benefit from well-licensed datasets, new tools and customized services to pursue their professional and personal goals.

The tasks within the use case can be organized according to three main areas:

- Making the Wolters Kluwer data available in a machine-readable form and then executing the interlinking and data enrichment tools of the LOD2 Stack on it.
- Creating a semantic knowledge layer based on this data and executing the editorial part of data management as well as general data visualization tools of the LOD2 Stack on it.
- Describing in more detail the business impact of this new kind of data in the media and publishing industry, especially with respect to expected hurdles in usage like governance and licensing issues.

1.2 Wolters Kluwer Company Profile

Wolters Kluwer Germany (WKD) is an information services company specializing in the legal, business and tax sectors. Wolters Kluwer provides pertinent information to professionals in the form of literature, software and services. Headquartered in Cologne, it has over 1,200 employees located at over 20 offices throughout Germany, and has been conducting business on the German market for over 25 years.

Wolters Kluwer Germany is part of the leading international information services company, Wolters Kluwer n.v., located in Alphen aan den Rijn (The Netherlands). The core market segments, targeting an audience of professional users, are legal, business, tax, accounting, corporate and finance services, and healthcare. Its shares are quoted on the Euronext Amsterdam (WKL), and are included in the AEX and the Euronext 100 indices. Wolters Kluwer has annual sales of €3.56 billion (2013), employs approximately 19,000 people worldwide and operates in over 40 countries throughout Europe, North America, the Asia Pacific region and in Latin America.

[3] For more detailed information see [5].

1.3 Data Transformation, Interlinking and Enrichment

This task has two main goals. The first goal is to adopt and deploy the LOD2 Stack[4] to the datasets of Wolters Kluwer. These datasets cover all document types being normally used in legal publishing (laws and regulations, court decisions, legal commentary, handbooks and journals). The documents cover all main legal fields of law like labor law, criminal law, construction law, administration law, tax law, etc. The datasets also cover existing legal taxonomies and thesauri, covering each a specific field of law, e.g. labor law, family law or social law. The overall amount of data (e.g. 600.000 court decisions) is large enough to make sure that the realistic operational tasks of a publisher can be executed with the data format and tools developed within the LOD2 project to support the respective use case. The datasets were analyzed according to various dimensions, e.g. actors, origin, geographical coverage, temporal coverage, type of data etc. relevant to the domain of legal information. Within the LOD2 project, all datasets were made available in formats adhering to open-standards, in particular RDF and Linked Data. Note that the datasets already existed in XML format at the start of the project and were transformed to RDF via XSLT script. The second goal is to automatically interlink and semantically enrich the Wolters Kluwer datasets. In order to achieve this, we leveraged the results from the LOD2 research work packages 3 and 4 (Chaps. 3 and 4) on the automated merging and linking of related concepts defined according to different ontologies using proprietary and open tools. Data from external sources (German National Library[5], DBpedia, STW[6], TheSoz[7] & EuroVoc[8]) were used to enrich the Wolters Kluwer datasets to leverage their semantic connectivity and expressiveness beyond the state of the art. This effort resulted in operational improvements at Wolters Kluwer Germany (WKG) as well as added-value for WKG customers. WKG was expecting that high current internal (manual) efforts concerning taxonomy and thesaurus development and maintenance would partly be substituted by integrating external LOD sources. This held also true for specific metadata like geographical information, detailed information about organizations and typical legal content itself from public issuing bodies. The internal workflow at WKG would therefore be enhanced as automated alerting (e.g. push notifications, see Chap. 5) could be executed to inform internal editors of data changes based on the underlying interlinked data. This would be a major gain as this process is currently labor intensive as it requires editors to physically monitor changes to content.

1.4 Editorial Data Interfaces and Visualization Tools

This task provided the main functionality for publishing, searching, browsing and exploring interlinked legal information. This included querying and facet-based

[4] See Chap. 6 and http://stack.linkeddata.org/, accessed May 10, 2014.

[5] http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html, accessed June 10, 2014.

[6] http://zbw.eu/stw/versions/latest/about, accessed June 10, 2014.

[7] http://www.gesis.org/en/services/research/thesauri-und-klassifikationen/social-science-thesaurus/, accessed June 10, 2014.

[8] http://eurovoc.europa.eu/, accessed June 10, 2014.

browsing of dataset metadata along various dimensions (dataset type, spatial/temporal coverage, origin etc.), as well as authoring of new metadata and data. It also investigated issues, such as access control and user rights management, to enable customized access levels for various user roles and clients. Additionally, different visualizations of e.g. geo location and statistical information were implemented (see LOD2 work package 5, Chap. 5).

1.5 Business Impact and Relevant Pre-conditions for Success

This task investigated into Intellectual Property Rights (IPR) management (licensing and management of usage rights) as well as business value of interoperable metadata. While traditional regimes, especially in the private sector, mostly rely on a Strong-IPR philosophy, by which the use and commercial exploitation of metadata is strictly regulated and governed, interoperable metadata requires more flexible licensing arrangements that take advantage of openness- and commons-based approaches. While the continuum between and the interplay of strong and light intellectual property rights for interoperable metadata is still a young and unexplored research field, it is the licensing strategy that defines the legal framework in which asset diversification and value creation takes place. The application of the uniform data model of RDF to metadata enables syntactic and semantic interoperability and leverages the network characteristics of metadata. While the lack of a uniform data model leads to proprietary lock-ins with respect to metadata assets like schemata, vocabularies, ontologies, indices, queries etc., interoperable metadata transcend these boundaries and open up possibilities for asset creation under the circumstances of economies of scale and positive feedback (Metcalfe's Law) as well as the social dynamics behind it (Reed's Law). Diversification for interoperable metadata can be looked at from a resource-based and a market-based point of view. The resource-based approach investigates how economically valuable resources are created and commercially exploited. The market-based approach looks at new customers and market segments that can be entered and secured.

2 Processing Data

The core challenge in this use case was to develop the (legal) data ecosystem by using the tools from the LOD2 Stack. Since the whole Semantic Web paradigm was new to WKD, we chose an iterative approach to learn and to optimize and smoothen the workflows and processes that come with it [4].

In order to focus on the highlights, we will not report on this iterative part here, but more on the results of every task. First, we built a knowledge framework based on the information we already stored in the XML documents. This led to an initial version of the knowledge graph describing our domain. We then executed LOD2 Stack tools [1] on this graph in order to enrich this information using data extraction technologies as well as executing data curation for cleansing; and linking tools for ingesting knowledge from external sources. Finally, we added a visualization layer (i) to support the editorial team in metadata management and (ii) to help our customers with visualizations supporting data analytics capabilities (see also [8]).

2.1 Transformation from XML to RDF

One major goal of the "Media and publishing" use case was to develop a stable transformation process for the WKG XML data. The development of the mapping schema from XML to RDF was based on the provided WKG DTD – so that the ontology was chosen to express the WKG data. The development of the schema for the transformation has been done in the following steps:

- Define vocabularies used for the WKD RDF schema (see Table 1)
- Define the URI pattern used for the WKD RDF schema
- Mapping definition
- Develop the XSLT style sheet based on the vocabularies and the URI patterns

In addition, a WKD schema description (http://schema.wolterskluwer.de) was developed, extending the used vocabularies to cover specific classes and properties. For the transformation of the WKD XML data to RDF various URI patterns had to be developed to cover the various types of data/information created:

- Resources (The transformed documents and document parts themselves)
  e.g. labor protection law http://resource.wolterskluwer.de/legislation/bd_arbschg
- Vocabularies (used to harmonize parts of the used metadata e.g. taxonomies, authors, organizations, etc.)
  e.g. labor law thesaurus http://vocabulary.wolterskluwer.de/kwd/Arbeitsschutz
- WKD Schema Vocabulary (Specific properties defined for the mapping schema)
  e.g. keyword http://schema.wolterskluwer.de/Keyword

The mappings between the WKD DTD and the WKD schema were implemented as XSLT functions. The WKD XML data was then transformed into RDF triples by applying the functions related to the relevant XML elements. Note that the output of the transformation was using the RDF/XML serialization.

Table 1. Schemas that have been evaluated and are used in the WKD RDF schema (applied vocabularies)

Vocabulary	Prefix	Namespace
BIBO	bibo	http://purl.org/ontology/bibo/
Dublin core	dc	http://purl.org/dc/elements/1.1/
Dublin core terms	dcterms	http://purl.org/dc/terms/
FOAF	foaf	http://xmlns.com/foaf/0.1/
Metalex	metalex	http://www.metalex.eu/metalex/2008-05-02#
OWL	owl	http://www.w3.org/2002/07/owl#
RDF	rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
RDF schema	rdfs	http://www.w3.org/2000/01/rdf-schema#
SKOS	skos	http://www.w3.org/2004/02/skos/core#
XHTML vocabulary	xhtml	http://www.w3.org/1999/xhtml/vocab#

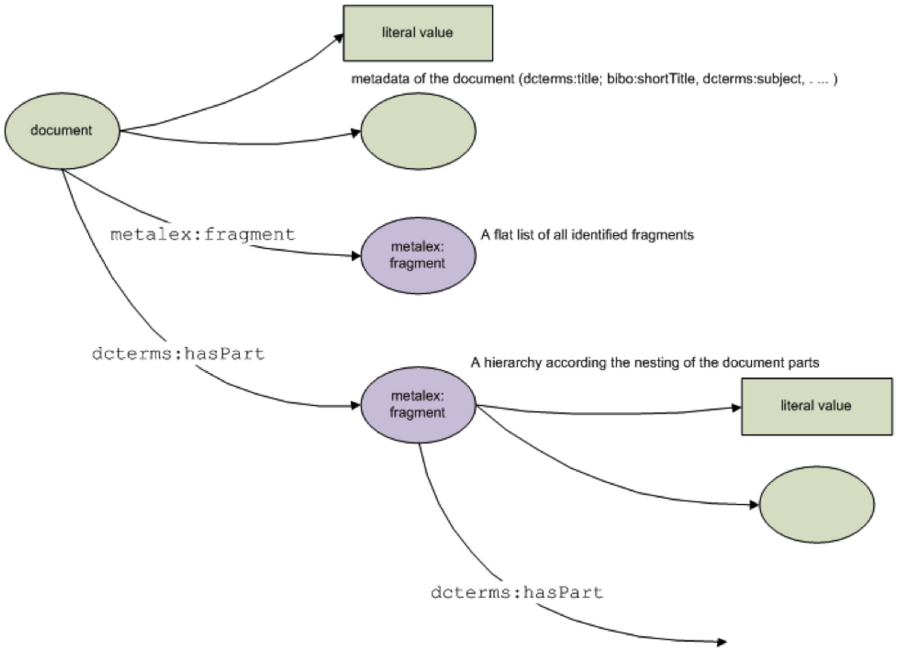

Fig. 1. RDF graph for a document URI

The transformation resulted in a number of triples, stored in a named graph per document (see Fig. 1). In this way, a provenance relationship between the existence of the triple, the XSLT template and the original XML document was created. If either the XSLT template or the XML document was updated, then the set of triples to be updated was uniquely identified with the graph name.

Valiant[9], a command line processing tool written in JAVA supporting XSLT2.0, has been developed for the transformation process within the work package . As a first step, Virtuoso Sponger Cartridge was explored, as Virtuoso[10] was part of the LOD2 Stack, but this track was abandoned due to the lack of support for XSLT 2.0. For the management of the taxonomies and vocabularies PoolParty[11] was used. Additionally, Venrich was developed to support the batch process for the alignment of the document metadata and the vocabularies and taxonomies. All the data was stored in Virtuoso.

The initial transformation resulted in:

- 785,959 documents transformed to RDF graphs with a total of 46,651,884 triples
- several taxonomies and vocabularies that have been created based on the data

[9] https://github.com/bertvannuffelen/valiant, accessed June 10, 2014.
[10] http://virtuoso.openlinksw.com/, accessed June 10, 2014.
[11] http://www.poolparty.biz/, accessed June 10, 2014.

Additionally, two of the developed vocabularies have been released as linked open data under an open source license by WKG[12,13].

2.2 Metadata Management Process

In the past, publishers have focused their content management systems around printed products: books, leaflets, journals, etc. A document centric approach, in which metadata and content are combined in one document, is well suited. Electronic publishing offers new opportunities, but also provides challenges for existing content management systems. For instance, it has changed the way people find information: instead of following the imposed structure and taking advantage of the printed index and the footnote system, electronic publishing allows jumping arbitrary through the publication following more closely a person's processes of thought. Without quality metadata this is unrealizable.

Having quality data is crucial for a publisher's business. Incomplete, erroneous or inaccurate information reduce the customers trust in the data, and hence in the publishing body. Consequently a large amount of effort in this work package was around improving and controlling the quality of the data. We will elaborate how the Linked Data representation of the extracted metadata is an enabler in the data quality processes.

The editorial process of a publisher like Wolters Kluwer Germany is today driven by 3 key stakeholders:

- The content editor creates the content: comments on law or jurisdictions, news, etc. Often the content editor is not part of the publisher organization, but an expert in the field who is using the publisher's dissemination channels to reach its audience. In the Use case contents are also partly harvested from legal institutions.
- The metadata editor manages the metadata of the content provided by the content editor.
- The taxonomist is responsible for the coherency and completeness of the controlled vocabularies used to create the metadata.

While applying the Linked Data paradigm on the editorial process a fourth role has emerged:

- the enrichment manager is a role which naturally emerges from the Linked Data paradigm. She is responsible for selecting external data sources that are thrust worthy and which contain data that provides added value to the content.

These stakeholders interact with each other via the content management system of the publisher (Fig. 2). The prototypical interaction pattern is the following. The content editor uploads a (new) version of a document. Via automated extractions, metadata is added. Inside the publishers organization the metadata editor is validating and

[12] See http://vocabulary.wolterskluwer.de/, accessed June 10, 2014.

[13] See further information about this in 3 Licensing Semantic Metadata and Deliverable 7.1.1 http://static.lod2.eu/Deliverables/Deliverable-7.1.1.pdf.

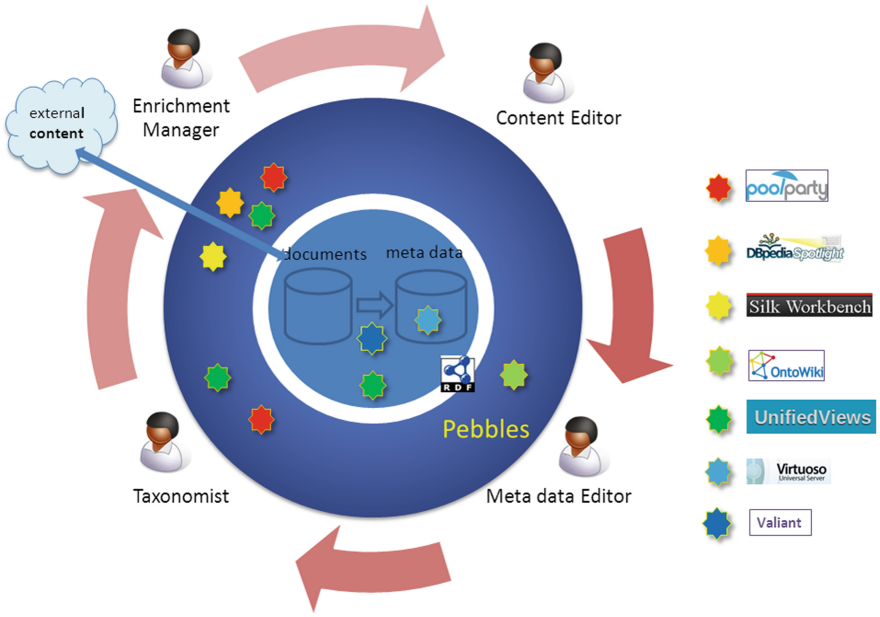

Fig. 2. Metadata management workflow

augmenting the attached metadata to make sure the document is ready for publication. In that process the metadata editor is using the controlled vocabularies that the taxonomist is maintaining. Controlled vocabularies need constant curation (responsibility of the taxonomist) in order to meet the ever changing world.

To explore how Linked Data can transform and support the metadata quality management and editorial process, a dedicated LOD2 Stack instance was setup. Since metadata quality is the center of the problem statement, the central component is formed by an adapted version of Ontowiki[14], called Pebbles. Pebbles supports the editing of the metadata independently of the content, and it is aimed for the metadata editors. For the taxonomists, software support is given by the PoolParty suite. And finally the enrichment manager is supported by a whole arsenal of tools of which Silk[15] and LOD Management Suite[16] – also called UnifiedViews (with the automated annotation processes for DBpedia Spotlight[17] and PoolParty Extractor[18]) - are the most notable.

Pebbles, a Metadata Editor

The user is welcomed in Pebbles with a dashboard overview showing the most recent updated documents and the documents with the most outstanding issues.

[14] http://aksw.org/Projects/OntoWiki.html, accessed June 10, 2014.

[15] http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/, accessed June 10, 2014.

[16] https://grips.semantic-web.at/display/public/LDM/Introduction, accessed June 10, 2014.

[17] https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki, accessed June 10, 2014.

[18] http://www.poolparty.biz/portfolio-item/poolparty-extractor/, accessed June 10, 2014.

Such dashboard view aids to focus on the most important items, but also it reflects the users' current state of work. After the user selects a document, Pebbles shows the document view on which the textual document content is shown together with its metadata. It is important that the metadata editor sees the document content in order to being able to validate the correctness of the associated metadata. Here the metadata can be updated, but also new metadata properties can be added according to the WKD schema. New properties can be added by hand, or result from the suggestions that are associated with the document (Fig. 3).

A suggestion is an association of some value with the document that has been added via an external process. This external process is controlled by the enrichment manager. The enrichment manager uses a linking environment (e.g. Silk) or an annotation engine (e.g. DBpedia Spotlight) to create these associations. At this point in time the enrichment manager has two options: either she directly adds the resulting associations to the metadata store, or the associations are reviewed through the quality assurance process. The quality assurance process is performed by the metadata editor by accepting/rejecting suggestions in the Pebbles environment. As the metadata editor that has the ownership of the documents metadata, she is the right person to make that decision. In case of the acceptance of a concept, the associated default metadata property can also be updated. This creates flexibility in the process: the enrichment manager can suggest new associations without deciding upfront the property which is handy in the case an annotation engine is being used. Such annotation engines often return related concepts belonging to a wide variety of domains (persons, countries, laws, …) It is however advised for the smoothness of the process to make the properties as concrete as possible.

The provided collection of documents by Wolters Kluwer forms an interconnected network. Journal articles refer to laws and court cases, and so on. In a document centric environment, these links are typically stored inside the document container. It is easy given a document to follow the outgoing references, whereas the reverse search (i.e. finding all documents that refer the current document) is much harder. Applying this

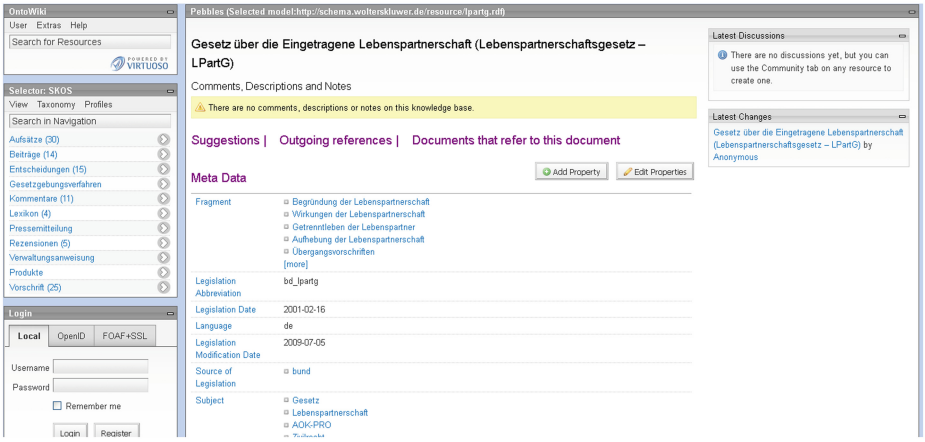

Fig. 3. Pebbles document metadata view

search on data in a RDF store simply requires inverting patterns in the SPARQL query. The non-directionality of the RDF graph model allows creating quickly any exploration path that is desired. Often exploration paths are quite generic: for instance to show the list of documents that belong to a particular document category is very similar to showing the list of documents for an author. By configuring the tree navigation widget with the values for a taxonomy, Pebbles offers a faceted approach to explore documents. The usability of the faceted browsing is determined by the quality of the taxonomies being used and the quality of the metadata that are tagging the documents.

Issues Identified in Existing Metadata, Based on Developed Vocabularies
During transformation of WKD data to RDF, several metadata properties were defined as having skos:Concepts as their range. The rationale behind that was that this data may be organized and managed in a next step in taxonomies or thesauri. In a second iteration after processing all the data, missing concepts have been detected and were added to the vocabularies.

During the review of the generated data, besides missing mappings to taxonomies the following issues in the existing metadata transformed to RDF were found:

- Transformation errors (e.g. concept generated with "" labels): To avoid this, the schema transformation has to be adapted to ignore empty metadata entries.
- Wrong Metadata (e.g. job titles or page numbers instead of organization name concerning the organizations taxonomy): This needs to be cleaned up manually. Rules can be provided to detect such kind of data during transformation; and the same rules could be applied to exclude this data from display in the metadata editor (Pebbles). Since this data can also be managed (changed/edited/deleted) in Pebbles, no additional efforts for a rule based cleaning have been made.
- Same concepts with different label: We decided that automatic mapping of metadata representing the same concepts (e.g. different spelling for persons, see Table 2 for different reasons) could not be done during schema transformation, because no quality assurance could be provided that way. So an interface for disambiguation of concepts based on label similarity was developed to provide a semi-automatic way of cleaning up those concepts.

Notification Service
We developed a scenario, where several vocabularies were developed and partly published as Linked Open Data (labor law thesaurus and courts thesaurus) with PoolParty. Furthermore, Pebbles was developed as an environment designed to manage RDF metadata for the WKD documents. To stay up-to-date with the latest changes in these datasets, the resource subscription and notification service (rsine[19], published under an open-source license at GitHub) was developed, allowing dataset curators to subscribe for specific changes that they are interested in and to get a notification as soon as such changes occur.

[19] https://github.com/rsine/rsine, accessed June 10, 2014.

Table 2. Possible issues for different author names

Confusions	First version	Second version	Third version
Family name change after marriage	Gritt Diercks	Gritt Dierks-Oppler	–
	Andrea Banse	Andrea Schnellbacher geb. Banse	Andrea Schnellbacher
Second forename	Bernd Schneider	Bernd Peter Schneider	–
Initials	Detlev Müllerhoff	D.Müllerhoff	–
Typos	Cornelius Prittwitz	Cornelins Prittwitz	–
Punctuation	Hans-Dieter Becker	Hans Dieter Becker	–
Different writings	Detlev Burhoff	Detlef Burhoff	–
Different characters	Østerborg	Österborg	Osterborg

Rsine is a service that tracks RDF triple changes in a triple store and creates a history of changes in a standardized format by using the change set ontology[20]. Users wanting to receive notifications can express the kind of changes they are interested in via SPARQL queries. These queries are sent to rsine, encapsulated in a subscription document that can also contain further information such as how the notification message should be formatted. Notifications were sent via mail.

The implemented scenarios focused on the following three main use cases:

- vocabulary management
- vocabulary quality
- metadata management

For all main use cases, several scenarios[21] have been implemented.

2.3 Enrichment of WKD Data

In a first step, the enrichment of WKD Data has been applied to the vocabularies published by WKD. The WKD Arbeitsrechtsthesaurus (labor law thesaurus) was linked (skos:exactMatch) with DBpedia[22], STW[23], Thesoz[24] and Eurovoc[25]. The WKD

[20] http://vocab.org/changeset/schema.html, accessed June 10, 2014.

[21] Scenarios are listed in Deliverable 5.3.2.

[22] http://de.dbpedia.org/, accessed June 10, 2014.

[23] Thesaurus for economics of the Leibniz Information Centre for Economics http://zbw.eu/stw/versions/latest/about, accessed June 10, 2014.

[24] Social science thesaurus of the Leibniz Institute of Social Sciences, http://www.gesis.org/en/services/research/thesauri-und-klassifikationen/social-science-thesaurus/, accessed April 18, 2014.

[25] Multilingual thesaurus of the European Union, http://eurovoc.europa.eu/, accessed June 15, 2014.

Table 3. Data added to the concepts of the WKD vocabulary

		DBpedia	STW	TheSoz	EuroVoc
skos:altLabel	Alternative wording		X	X	X
skos:scopeNote	Note		X	X	X
dbpedia-owl:abstract	Abstract/short definition	X			
dcterms:subject	Subject	X			
rdfs:label	Wording	X			
foaf:page	Related pages	X			
dbpedia-owl: thumbnail	Picture	X			
geo:long	Longitude	X			
geo:lat	Latitude	X			

Gerichtsthesaurus (courts thesaurus) was linked to DBpedia. In addition to linking to the respective sources, the WKD vocabularies have been enriched by including data from the respective sources (see Table 3). The provenance of the included data has been preserved, storing the added data in separate graphs.

The mapping to those resources was based on the similarity of concept labels and has been done in a semi-automatic process using Silk. Figure 4 shows the evaluation workspace where the user can check and accept or decline the suggested links. The enrichment with additional data as shown in Table 3 has been done automatically using the LOD Management Suite.

The published vocabularies are publicly available under CC BY 3.0. The frontend uses the data from the external datasets to enhance the user experience. For instance, the geographic location of courts can be leveraged to be displayed on a map (Fig. 5). The map is also available on the detail pages of the courts, where images, showing from DBpedia are also displayed, showing mostly the court building.

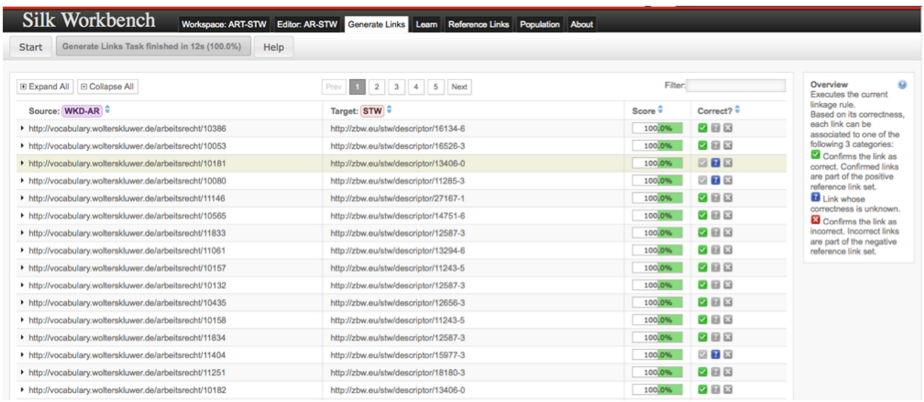

Fig. 4. Mapping results in Silk

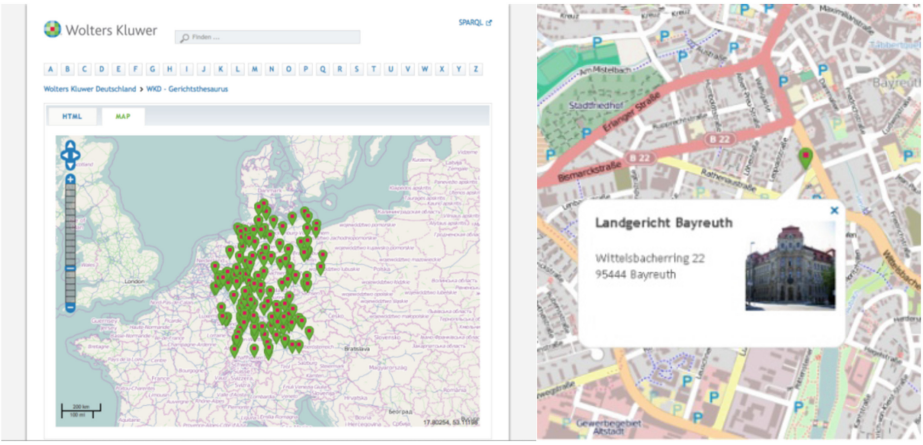

Fig. 5. Court maps of courts within Germany and a specific court

In a second step, WKD document data has been enriched by linking to external resources. Legislations were matched (skos:exactMatch) with DBpedia information – data coming from Wikipedia info boxes (especially scope and practice area) could be used to enrich the documents further. Authors were linked to persons in the GND[26] dataset (Integrated Authority File of the German National Library) – links to these external sources are included in Pebbles. The GND dataset contains more metadata about authors than WKD does collect. Both mappings to DBpedia and GND were done using Silk (Table 4). A third enrichment project took place with the EU Publication Office to match documents from the EU Cellar platform with documents from Pebbles.

Entity Extraction was another approach to enrich the metadata of documents. It was tested randomly with two tools: DBpedia Spotlight and the PoolParty Extractor.

Table 4. Overview of WKD concept linking to external and internal sources

Links to external/internal sources	Links
Courts thesaurus to DBpedia	997
Extended version of courts thesaurus	–
Labor law thesaurus to DBpedia	776
Labor law thesaurus to Thesoz	443
Labor law thesaurus to STW	289
Labor law thesaurus to EuroVoc	247
Legislations to DBpedia	155
Authors to GND	941
WKD Labor Law Thesaurus to WKD subjects	70

[26] http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html, accessed June 10, 2014.

Spotlight uses DBpedia concepts as an extraction base, whereas PoolParty uses the predefined controlled vocabularies. Both solutions provided a good overview on the topics and main issues of the tested documents. Nonetheless, the main problem of ambiguity appeared in both approaches and resulted in terms that came from different contexts (e.g. "deadline" that could mean the application deadline or the deadline of a labor agreement and therefore address different areas of labor law).

2.4 Visualization

Special features and visualization functionalities are crucial as part of the activities related to the publishing content supply chain. Visualizations are not only created for the internal management of data, but also for enabling product developments for the customers of information services of WKD. Therefore, we investigated options for presenting the created data in an appealing manner.

The visualization of controlled vocabularies provides different interfaces depending on the dataset. For instance, courts are visualized in form of a map in the Linked Data frontend[27], where the geographical information is used either to visualize courts as pins on a map of Germany or a local map presenting the geolocation information for each individual court (see Fig. 5).

For the labor law thesaurus we chose the visualization in form of a semantic network. Concepts are shown with related concepts within the same context of labor law (Fig. 6).

Fig. 6. Semantic net of labor law

[27] At http://lod2.wolterskluwer.de/, accessed May 10, 2014.

The visualization of overall datasets is possible with CubeViz[28] and gives an insight to the amount of available data for specific document types, practice areas, time frames and courts (Fig. 7).

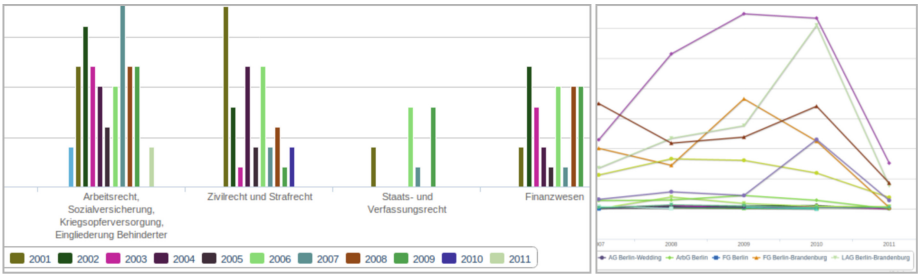

Fig. 7. Laws of specific practice areas per year; jurisdictions per court per year

Visualizing the data has proven to be an important step to give the user deeper understanding of the underlying data and to provide contextual information that can give new insights.

3 Licensing Semantic Metadata

Data is not an easy subject to talk about, especially when doing it from an economic perspective. From all intangible assets imaginable, data is a highly controversial one, given that its economic characteristics are hard to describe and even more difficult to protect. But to own, to control, and to share data one needs to define policies that describe the conditions under which data can be (re-)used in various contexts and for various purposes. Licensing is one such policy that allows us to define data as an economic good. Accordingly, data licensing is crucial in the development of data-driven businesses as it defines the properties of data in a generic economic sense in the dichotomies of scarcity-abundance, private-public and rivaling-complementary. Licenses are an enabler and a barrier for economic transactions. They set the boundaries in which economic actions take place and they define the legitimate or illegitimate usage of data for commercial or non-commercial purposes.

Beside licensing, technology as a constructivist framework for the creation and utilization of data plays an important role. Technology defines the good characteristics of data. According to this premise, it makes a difference whether data is generated manually or algorithmically; or optimized for syndication or storage within a silo etc. Technology influences the context in which data is being generated and utilized, thus changing the hermeneutic conditions under which data is being defined. It makes a difference, whether data is being treated as a solitary thing or linked for purposes of knowledge discovery and targeted insights. Hence it is crucial to gain a good

[28] http://aksw.org/Projects/CubeViz.html, accessed June 10, 2014.

understanding of the technology with which data has been generated to make economic sense out of it.

3.1 Traditional Protection Instruments for Intellectual Property

Semantic metadata is a fairly new kind of intellectual asset that is still subject to debate – concerning the adequate protection instruments [12]. Table 5 gives an overview on the applicability of various protection instruments. The table illustrates the complex nature of semantic metadata as intellectual property. Various instruments can be applied to various assets; while copyright, database right and competition right are the most relevant ones.

Copyright basically protects the creative and original nature of a literary work and gives its holder the exclusive legal right to reproduce, publish, sell, or distribute the matter and form of the work. Hence, any literary work that can claim a sufficient degree of originality can be protected by copyright.

Database Right protects a collection of independent works, data or other materials, which have been created with considerable financial investment, are arranged in a systematic or methodological way and are individually accessible by electronic or other means. Databases are also protected as literary works and need to have a sufficient degree of originality that requires a substantial amount of investment.

An Unfair Practices Act protects rights holders against certain trade practices, which are considered unfair in terms of misappropriation, advertising, sales pricing or damages to reputation. Especially the first aspect is relevant to semantic metadata, which actually occurs, when data is being reused without appropriate compensation i.e. in terms of attribution or financial return.

Patenting protects the inventory aspects of a novel technical artefact. Hence it does not directly impact the protection of semantic metadata as – at least in Europe – patents can just be acquired for hardware-related inventions. But as soon as semantic metadata becomes an indispensable subject of a methodology that generates physical effects, has a sufficient level of inventiveness and can be exploited commercially, these components can be protected under Patent Law.

Table 5. IPR instruments for semantic metadata [9]

	Copyright	Database right	Unfair practice	Patents
Documents	YES	YES	YES	NO
Dataset	NO	YES	PARTLY	NO
Description	YES	NO	YES	NO
Identifier	NO	NO	NO	NO
Name space	YES	YES	YES	NO
Vocabulary	PARTLY	YES	YES	NO
Classification	PARTLY	PARTLY	PARTLY	NO
Ontology	PARTLY	YES	YES	PARTLY
Rules	PARTLY	YES	YES	PARTLY

This overview conceals the fact that there exist regional differences in the practical application of IPR instruments. These differences and specificities of so called IPR regimes make the licensing of Linked Data a complex and sometimes confusing issue. I.e. while in the USA data is generally protected under the US copyright law[29], the European Union additionally provides the instrument of Database Right[30] to fill certain gaps between the various national copyrights of the EU member states. Additionally while the US Patent Act[31] allows the patenting of software, which also includes collections of data as output of an algorithmic process; this is formally forbidden in Europe under Article 52 of the European Patent Convention[32].

This situation has long been scrutinized by critics of traditional IPR practices. On the one hand, the differences between the various regional regimes lead to judicial uncertainty. On the other hand, the overlapping and complementary protection instruments tend to favor an "overprotection" of intellectual assets that stifle competition and innovation and prevent the development of business models and new ways of value creation (i.e. [2, 3, 6, 11]).

As a reaction to these structural externalities of the traditional IPR system, new licensing instruments have emerged over the past few years that deliberately focus on the creative and self-governed re-purposing of intellectual property with the aim to foster innovation, collaborative creation of value and finally the public domain. These so called commons-based instruments – well known under Creative Commons and lately Open Data Commons – play an important role in the commercial and non-commercial appropriation of Linked Data and are an important part of a Linked Data licensing policy. Additionally, we will discuss the purpose of so called "community norms" as a third important component in Linked Data licensing policy.

3.2 Licensing Policies for Linked Data

The open and non-proprietary nature of Linked Data design principles allow to easily share and reuse data for collaborative purposes. This also offers new opportunities for data publishers to diversify their assets and nurture new forms of value creation (i.e. by extending the production environment to open or closed collaborative settings) or unlock new revenue channels (i.e. by establishing highly customizable data syndication services on top of fine granular accounting services based on SPARQL).

To meet these requirements, commons-based licensing approaches like Creative Commons[33] or Open Data Commons[34] have gained popularity over the last few years, allowing maximum re-usability while providing a framework for protection against unfair usage practices and rights infringements. Nevertheless, to meet the requirements

[29] See http://www.copyright.gov/title17/, accessed July 10, 2013.

[30] See http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML, accessed July 10, 2013.

[31] See http://www.law.cornell.edu/patent/patent.overview.html, accessed July 10, 2013.

[32] See http://www.epo.org/law-practice/legal-texts/html/epc/2010/e/ma1.html, accessed July 10, 2013.

[33] See http://creativecommons.org/, visited April 22, 2012.

[34] See http://opendatacommons.org/, visited April 22, 2012.

of the various asset types, a Linked Data licensing strategy should make a deliberate distinction between the database and the content stored in it (see Table 5). This is necessary as content and databases are distinct subjects of protection in intellectual property law and therefore require different treatment and protection instruments. An appropriate commons-based protection strategy for a data provider could look as follows:

The contents of a linked dataset, which are comprised of the terms, definitions and its ontological structure, are protected by a CC-By v3.0 License[35], which allows the commercial and non-commercial reuse of any published artefact as long as the owner is mentioned.

The underlying database, which is comprised of all independent elements and works that are arranged in a systematic or methodological way and are accessible by electronic or other means, are protected by a ODC-By v1.0 License[36], which also allows the commercial and non-commercial reuse of any published artefact as long as the owner is mentioned.

Additionally to these two aspects, the licensing strategy also should incorporate a Linking Policy Community Norm, which explicitly defines the expectations of the rights holder towards good conduct when links are made to the various artefacts provided in the dataset.[37] This norm should provide administrative information (i.e. creator, publisher, license and rights); structural information about the dataset (i.e. version number, quantity of attributes, types of relations) and recommendations for interlinking (i.e. preferred vocabulary to secure semantic consistency).

All in all the three elements of a commons-based licensing policy – the CC-By v3.0 License, the ODC-By v1.0 License and the Community Norm – provide a secure and resilient judicial framework to protect against the unfair appropriation of open datasets.

3.3 Rights Expression Languages for Linked Data Licenses

The basic idea of Linked Data is to create an environment where information can flow freely and can be repurposed in multiple ways, not necessarily evident at the time of content creation. This holds true for open and closed settings alike. Hence a clear machine-readable explication of prohibits and permits associated with the usage rights of linked datasets is a necessary precondition to realize the promises of the Linked Data vision.

Open Digital Rights Language (ODRL): With the emergence and mass adoption of Digital Rights Management Systems since the end of the 1990s, several attempts have taken place to define machine-readable standards for the expression of rights over digital assets. One of these endeavors was ODRL, an XML vocabulary to express rights, rules, and conditions – including permissions, prohibitions, obligations, and

[35] See http://creativecommons.org/licenses/by/3.0/, visited April 20, 2012.

[36] See http://opendatacommons.org/category/odc-by/, visited April 20, 2012.

[37] See for example the community norm provided by the Leibniz Information Centre for Economics: http://zbw.eu/stw/versions/8.08/mapping/gnd/, accessed April 20, 2012.

assertions – for interacting with online content.[38] The corresponding ODRL Standards Group[39], a member of the World Wide Web (W3C) Community and Business Groups[40] since 2011, acts as an international initiative to define the specifications for expressing policy information over digital content residing on the Open Web Platform (OWP)[41].

ODRL utilizes an Entity-Attribute-Value Model to express a policy about rights and restrictions associated with a digital artefact. The legal information about allowed actions with a media asset (i.e. copying, sharing, modifying, attributing etc.) can be expressed within the ODRL vocabulary. Hence ODRL basically provides a machine-readable policy framework that supports the flexible and fine-granular definition of usage rights within dynamic usage settings like the web and other multi-channel environments. In 2013, the International Press and Telecommunications Council (IPTC) adopted ODRL as the basis for its Rights Markup Language (RightsML)[42,43]. Still in an experimental phase, the RightsML has mainly been applied to specify rights and restrictions with respect to photos[44], but its application goes beyond this specific asset type.

Besides ODRL, the Creative Commons Community has developed *Creative Commons Rights Expression Language*[45] (CCREL) to represent the various CC licenses in a machine-readable format. CCREL is the product of an informal W3C working group that issued its specifications in 2008. Since then, CCREL is being recommended by the Creative Commons Foundation as a standard for the machine-readable provision of Creative Commons licensing information to the public. Although never acknowledged as an official W3C recommendation, CCREL has evolved into a de facto standard for the special domain of Creative Commons Licenses and is expected to spread with the increasing need to provide explicit licensing information for automated processing on the web.

CCREL basically complements the ODRL vocabulary. It provides a condensed and hierarchically ordered set of properties that define the actions allowed with certain licenses. These properties can be seamlessly integrated into the ODRL vocabulary and allow to define fine-grained usage policies and constraints associated with a certain asset that falls into the legal domain of Creative Commons.

Generally it is important to mention that a combination of ODRL and CCREL is not obligatory to provide machine-readable licensing information on the web. The semantic expressivity of CCREL is sufficient to simply annotate existing assets with licensing information for automated processing. But in case of very complex and

[38] A comparable endeavour to create a data model for machine-readable statements on IPR in e-commerce transactions can be traced back to the year 1999. For details see [10].

[39] http://www.w3.oeg/cumunity/odrl/, accessed June 17, 2013.

[40] http://www.w3.org/community/, accessed June 17, 2013.

[41] http://www.w3.org/wiki/Open_Web_Platform, accessed June 17, 2013.

[42] http://dev.iptc.org/RightsML, accessed June 17, 2013.

[43] http://dev.iptc.org/RightsML-Introduction-to-ODRL, accessed July 1, 2013.

[44] http://dev.iptc.org/RightsML-10-Implementation-Examples, accessed June 17, 2013.

[45] http://www.w3.org/Submission/CCREL/, accessed July 1, 2013.

differentiated usage scenarios, a combination of ODRL and CCREL will be necessary, as ODRL provides the necessary semantic expressivity to define fine-granular usage policies associated with a certain asset that goes beyond the simple explication of licensing information, i.e. for the purposes of Digital Rights Management.

Beside Creative Commons, which is basically an extension of copyright, the *Open Data Commons* initiative[46] has started to provide legal tools for the protection of commons-licensed data assets. This is necessary as diverging regional judicial regimes require different IPR instruments to fully protect the various assets involved in the digital processing of information. For instance, data sources are protected by copyright in the USA, while in the European Union the protection of data sources is additionally complemented by so called database rights as defined in the Database Directive (96/9/EC)[47]. Hence to fully protect datasets in the European Union, it is actually necessary to provide legal information on various asset types from which certain parts can be licensed under Creative Commons, while others require Open Data Commons.

In contrast to ODRL and CCREL, the Open Data Commons initiative has not yet provided a REL of its own and it is to question whether this is necessary as licenses of Open Data Commons can be easily integrated in the vocabulary of other RELs.

4 Conclusion

The "Media and Publishing" use case has shown – based on real requirements from a global information service provider – that the expected added value to legal products and company processes can be achieved when using Linked Data and the accompanying Semantic Web technologies.

As a major outcome of this project, some tools from the LOD2 Stack like PoolParty and Virtuoso are already implemented and used in the operational systems of WKD. In that sense, the LOD2 Stack has shown its value for enterprises even before the project terminated.

The steps taken, described in this chapter, are most likely representative for many use case scenarios where Linked Data comes into play. First, existing internal and external data must be transformed into standard formats like RDF and SKOS. Then tools need to be utilized for further enrichment and linking and the resulting enhanced domain knowledge network needs to be further maintained and its content translated into functionalities in products[48]. This also covers different areas of visualization, which we investigated. Finally, governance and licensing of data need to be properly addressed, which is still at the end of the project a major issue. Potential business impact could be shown, but when the data is not usable in professional environments, it will not be taken up in the end.

---

[46] http://opendatacommons.org/, accessed July 1, 2013.

[47] http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML, accessed July 4, 2013.

[48] https://www.jurion.de, accessed June 10, 2014.

However, the steps taken are not at all easy. Already during transformation, the paradigm shift from a content to a data centric point of view raised a lot of questions and issues around quality, especially around normalization and granularity of information. This included the generation and maintenance of valid and stable identifiers. This challenge continued during enrichment phase, where the importance of identifiable and comparable contexts became obvious in order to link things properly and not to compare apples and oranges. During visualization activities, an important aspect, which was new to us in its consequence, was the need for consistent and complete data, which is normally not available when coming from a content based approach. So actually, the process that we faced was not only a technical and data driven one, it also changed our mindset when looking at data and its importance for our future business. In this respect, an important aspect we were not able to cover in this chapter is that of new business models based on Linked Data. Detailed information will be available at the end of the project at the project website.

All the major cornerstones for success mentioned above need to be further elaborated in the future.

Wolters Kluwer will actively participate in further research projects to make more data – and more clean (!) data – publicly available; to add more sophisticated tools to the open source tool stack of LOD2; and to address the licensing challenge within a growing community of data providers and customers of this data. First conversations with public information providers (e.g. with the Publications Office of the European Union or the German National Library) indicate common interests across and beyond traditional company boundaries.

References

1. Auer, S., Bühmann, L., Dirschl, C., Erling, O., Hausenblas, M., Isele, R., Lehmann, J., Martin, M., Mendes, P.N., van Nuffelen, B., Stadler, C., Tramp, S., Williams, H.: Managing the life-cycle of linked data with the LOD2 stack. In: Cudré-Mauroux, P., Heflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Parreira, J.X., Hendler, J., Schreiber, G., Bernstein, A., Blomqvist, E. (eds.) ISWC 2012, Part II. LNCS, vol. 7650, pp. 1–16. Springer, Heidelberg (2012)
2. Barton, J.H.: Adapting the intellectual property system to new technologies. In: Wallerstein, M.B., Mogee, M.E., Schoen, R.A. (eds.) Global Dimensions of Intellectual Property Rights in Science and Technology, pp. 256–283. National Academic Press, Washington (1993)
3. Bessen, J., Meurer, M.: Patent Failure: How Judges Bureaucrats and Lawyers Put Innovators at Risk. Princeton University Press, Princeton (2008)
4. Dirschl, C., Eck, K., Lehmann, J., Bühmann, L., Auer, S.: Facilitating data-flows at a global publisher using the LOD2 stack. Submitted to the Semant. Web J.
5. Hondros, C.: Standardizing legal content with OWL and RDF. In: Wood, D. (ed.) Linking Enterprise Data, pp. 221–240. Springer, New York (2010)

6. Klemens, B.: Math You can't Use: Patents, Copyright and Software. Brookings Institution Press, Washington (2006)
7. Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C., Lee, R.: Media meets semantic web – how the BBC uses DBpedia and linked data to make connections. In: Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., Simperl, E. (eds.) ESWC 2009. LNCS, vol. 5554, pp. 723–737. Springer, Heidelberg (2009)
8. Lee, S., Kim, P., Seo, D., Kim, J., Lee, J., Jung, H., Dirschl, C.: Multi-faceted navigation of legal documents. In: 2011 International Conference on and 4th International Conference on Cyber, Physical and Social Computing (2011)
9. Pellegrini, T.: Linked data licensing – Datenlizenzierung unter netzökonomischen Bedingungen. In: Schweighofer, E., Kummer, F., Hötzendorfer, W. (Hg.) Transparenz. Tagungsband des 17. Internationalen Rechtsinformatik Symposium IRIS 2014. Wien: Verlag der Österreichischen Computergesellschaft, S. 159–168 (2014)
10. Rust, G., Bide, M.: The <indecs> metadata framework - principles, model and data dictionary. http://www.doi.org/topics/indecs/indecs_framework_2000.pdf (2000). Accessed 18 July 2013
11. Sedlmaier, Roman: Die Patentierbarkeit von Computerprogrammen und ihre Folgeprobleme. Herbert Utz Verlag, München (2004)
12. Sonntag, M.: Rechtsschutz für Ontologien. In: Schweighofer, E., Liebwald, D., Drachsler, M., Geist, A. (eds.) e-Staat und e-Wirtschaft aus rechtlicher Sicht, pp. 418–425. Richard Boorberg Verlag, Stuttgart (2006)