Semantic Web 1 (2013) 1–5
IOS Press

Facilitating Data-Flows at a Global Publisher using the LOD2 Stack

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Christian Dirschl [a], Katja Eck [a], Jens Lehmann [b,*], Lorenz Bühmann [b], Sören Auer [c]
[a] *Wolters Kluwer Deutschland GmbH, 85716 Unterschleissheim, Germany*
E-mail: CDirschl@wolterskluwer.de, KEck@wolterskluwer.de
[b] *University of Leipzig, Institute of Computer Science, AKSW Group, Augustusplatz 10, D-04009 Leipzig, Germany*
E-mail: {lastname}@informatik.uni-leipzig.de
[c] *University of Bonn, Computer Science, Enterprise Information Systems & Fraunhofer IAIS, Bonn, Germany*
E-mail: auer@cs.uni-bonn.de
This requires a transformation of the publishing workflows towards the production of much richer meta-data for fine-grained and highly interlinked pieces of content. Linked Data can play a crucial role in this transition. The LOD2 Stack is an integrated distribution of aligned tools which support the whole lifecycle of Linked Data from extraction, authoring/creation via enrichment, interlinking, fusing to maintenance. In this application paper, we describe a real-world usage scenario of the LOD2 stack at a global publishing company. We give an overview over the LOD2 Stack and the underlying life-cycle of Linked Data, describe data-flows and usage scenarios at a publisher and then show how the stack supports those scenarios.

Abstract. The publishing industry is at the verge of an era, wherein particular professional customers of publishing products are not so much interested in comprehensive books and journals, i.e. traditional publishing products, anymore as they now are interested in possibly structured information pieces delivered just-in-time as a certain information need arises. This requires a transformation of the publishing workflows towards the production of much richer meta-data for fine-grained and highly interlinked pieces of content. Linked Data can play a crucial role in this transition. The LOD2 Stack is an integrated distribution of aligned tools which support the whole lifecycle of Linked Data from extraction, authoring/creation via enrichment, interlinking, fusing to maintenance. In this application paper, we describe a real-world usage scenario of the LOD2 stack at a global publishing company. We give an overview over the LOD2 Stack and the underlying life-cycle of Linked Data, describe data-flows and usage scenarios at a publisher and then show how the stack supports those scenarios.

Keywords: Publishing, Linked Open Data, LOD2 Stack

1. Introduction

In times of tablets, smartphones and a growing number of other electronic devices, publishers are more and more forced to move towards electronic publishing. Possibilities for consuming information are changing and so do the expectations of customers. However, documentation and processing of publisher's data are not yet designed for such purposes. Many structural requirements like separation of XML and metadata or the inclusion of external knowledge sources are cur-

*Corresponding author. E-mail: lehmannn@informatik.uni-leipzig.de.

rently not implemented because they were not needed before. Now that new digital work environments offer new functionalities (e.g. non-linear story telling, inclusion of background information, delivery of just-in-time, context-specific and personalized information), internal workflow processes especially at publishers targeting professional audiences (e.g. legal, tax, accounting professionals) have to be adapted in order to generate best possible value.

Semantic technologies can help to support these processes. Publishers still deal with large amounts of unstructured, textual content. Knowledge extraction approaches can help to annotate and enrich such content. Once formalized knowledge (e.g. adhering to the RDF data model) is extracted it needs to be stored, managed and made available for querying. Links to other knowledge bases, either from the publisher itself, from other content providers or from the Web of Data, need to be established. We can apply reasoning techniques to enrich the knowledge bases with upper-level structures. The evolution of the extracted knowledge needs to be supported, since the original documents might change. Finally, semantic search, exploration and visualization techniques can help to gain new insights from the semantically represented content. The LOD2 Stack provides specialized tools for each of these lifecycle stages and can consequently be used to facilitate the semantic content processing workflows at a publisher.

Let's assume the following real world user scenario: Gerhard, an accounting professional works for TWC, a leading tax and accounting consultancy. He is responsible for certifying the value-added-tax (VAT) returns at the Europe-wide operating food retailer named Aldo (customer of the tax and accounting consultancy). For Gerhard it is crucial, to track all changes of VAT regulations in all the countries Aldo operates in, which might include countries in the Euro zone other EU member states and a few neighboring countries. As a result, Gerhard needs to be informed, whenever a law in one of these countries related to VAT regulations is changed. In addition, he has to track court decisions of all cases related to VAT regulations. Because Aldo has to update its ERP systems when VAT regulations change, Gerhard wants to notify Aldo's IT department already proactively as early as possible, when major regulatory changes (e.g. increase of the VAT for certain products in a certain country) are planned. Currently, Gerhard and his team has to track a vast number of textual sources provided to TWC by a global publisher and several smaller regional publishers specialized in the legal, accounting and tax domains for relevant legislation and regulatory changes. In future, the global publisher aims to deliver Gerhard and his colleagues at TWC much more personalized and context specific information pieces fulfilling exactly his information need. Gerhard aims to register the sources (legislation in certain countries, court decisions and parliamentary initiatives) he wants to track and filter information according to entries in a taxonomy related to VAT. Subsequently, Gerhard wants to be notified whenever a certain piece of information related to his particular information need is published by one of the identified sources. He wants to easily compare the changes applied to a certain law, for example, and be able to explore specifiably related court decisions.

Such a scenario requires an interaction and data exchange between different companies as well as several intelligent systems to process data as well as possibly enrich and interlink it. Single tools cannot solve those problems in isolation at large scale. Several interoperable components based on standards are required to achieve this. In this application report, we describe how the LOD2 stack can address those challenges.

This application report builds on [1]. While the previous article focused on a detailed characterisation of the LOD2 stack, this application report focuses on a detailed description of the usage scenarios at a global publisher.

The application report is structured as follows: In Section 2, we present the LOD2 stack architecture on a high level. After that, the vision of the vision of the Linked Data lifecycle is explained in Section 3. The phases of this lifecycle are closely related to the dataflows at global publishers, which are described in Section 4. Based on this, in Section 5, we describe how the LOD2 stack was applied at a particular publisher – Wolters Kluwer. We follow up with a brief summary of related work in Section 6 and plans for future work in Section 7.

2. Overview of the LOD2 Stack

The description of the LOD2 stack and the Linked Data lifecycle are extensions of earlier work in [1] and [3]. The LOD2 Stack is an integrated distribution of components, which support the whole lifecycle of Linked Data from extraction, authoring/creation via enrichment, interlinking, fusing to exploration. The major components of the LOD2 Stack are open-source in order to facilitate wide deployment. Through an it-

erative software development approach, the stack contributors aim at ensuring that the stack fulfills a broad set of user requirements and thus facilitates the transition to a Web of Data. The stack is designed to be versatile; for all functionality we define clear interfaces, which enable the plugging in of alternative third-party implementations.

In order to fulfill these requirements, the architecture of the LOD2 Stack is based on three pillars:

1. Software integration and deployment using the Debian packaging system: The Debian packaging system is one of the most widely used packaging and deployment infrastructures and facilitates packaging and integration as well as maintenance of dependencies between the various LOD2 Stack components. Using the Debian system also allows to facilitate the deployment of the LOD2 Stack on individual servers, cloud or virtualization infrastructures.
2. Use of a central SPARQL endpoint and standardized vocabularies for knowledge base access and integration between different tools: All components of the LOD2 Stack access this central knowledge base repository and write their findings back to it. In order for other tools to make sense out of the output of a certain component, it is important to define vocabularies for each stage of the Linked Data life-cycle.
3. Integration of the LOD2 Stack user interfaces based on REST enabled Web Applications: Currently, the user interfaces of the various tools are technologically and methodologically quite heterogeneous. We do not resolve this heterogeneity, since each tool's UI is specifically tailored for a certain purpose. Instead, we develop a common entry point for accessing the LOD2 Stack UI, which then forwards a user to a specific UI component provided by a certain tool in order to complete a certain task.

These three pillars comprise the methodological and technological framework for integrating the very heterogeneous LOD2 Stack components into a consistent framework.

3. The Linked Data Lifecycle

The different stages of the Linked Data life-cycle, depicted in Figure 1, include:

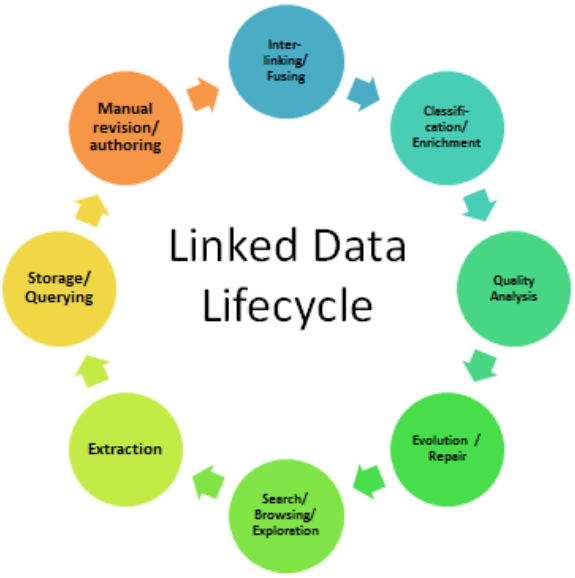

Fig. 1. Stages of the Linked Data life-cycle supported by the LOD2 Stack.

1. *Storage:* Efficient RDF data management techniques fulfilling requirements of global publishers comprise column-store technology, dynamic query optimization, adaptive caching of joins, optimized graph processing and cluster/cloud scalability.

1. *Authoring:* The LOD2 Stack facilitates the authoring of rich semantic knowledge bases, by leveraging Semantic Wiki technology, the WYSIWYM paradigm (What You See Is What You Mean) and distributed social, semantic collaboration and networking techniques.
2. *Interlinking:* Creating and maintaining links in a (semi-)automated fashion is still a major challenge and crucial for establishing coherence and facilitating data integration as outlined in the publishing usage scenario in the introduction. We seek linking approaches yielding high precision and recall, which configure themselves automatically or with end-user feedback.
3. *Classification:* Linked Data on the Web is mainly raw instance data. For data integration, fusion, search and many other applications, however, we need this raw instance data to be classified into taxonomies. In the LOD2 stack, semi-automatic components for this purpose are included.
4. *Quality:* The quality of content on the Data Web varies, as the quality of content on the document web varies. The LOD2 Stack comprises tech-

niques for assessing quality based on characteristics such as provenance, context, coverage or structure. The goal in our application scenarios is to assess whether data sources for a publisher are complete, consistent, reliable etc.

5. *Evolution/Repair:* Data on the Web is dynamic. We need to facilitate the evolution of data while keeping things stable. Changes and modifications to knowledge bases, vocabularies and ontologies should be transparent and observable. The LOD2 Stack comprises methods to spot problems in knowledge bases and to automatically suggest repair strategies.
6. *Search/Browsing/Exploration:* For many users, the Data Web is still invisible below the surface. LOD2 develops search, browsing, exploration and visualization techniques for different kinds of Linked Data (i.e. spatial, temporal, statistical), which make the Data Web sensible for real users.

These life-cycle stages, however, should not be tackled in isolation, but by investigating methods which facilitate a mutual fertilization of approaches developed to solve these challenges. Examples for such mutual fertilization between approaches include:

1. The detection of mappings on the schema level, for example, will directly affect instance level matching and vice versa.
2. Ontology schema mismatches between knowledge bases can be compensated for by learning which concepts of one are equivalent to which concepts of another knowledge base.
3. Feedback and input from end users (e.g. regarding instance or schema level mappings) can be taken as training input (i.e. as positive or negative examples) for machine learning techniques in order to perform inductive reasoning on larger knowledge bases, whose results can again be assessed by end users for iterative refinement.
4. Semantically enriched knowledge bases improve the detection of inconsistencies and modelling problems, which in turn results in benefits for interlinking, fusion, and classification.
5. The querying performance of RDF data management directly affects all other components, and the nature of queries issued by the components affects RDF data management.

As a result of such interdependence, we pursue with the LOD2 Stack the establishment of an improvement cycle for knowledge bases on the Data Web. The improvement of a knowledge base with regard to one aspect (e.g. a new alignment with another interlinking hub) triggers a number of possible further improvements (e.g. additional instance matches).

The challenge is to develop techniques which allow exploitation of these mutual fertilizations in the distributed medium Web of Data. One possibility is that various algorithms make use of shared vocabularies for publishing results of mapping, merging, repair or enrichment steps. After one service published its new findings in one of these commonly understood vocabularies, notification mechanisms (such as *Semantic Pingback* [12]) can notify relevant other services (which subscribed to updates for this particular data domain), or the original data publisher, that new improvement suggestions are available. Given proper management of provenance information, improvement suggestions can later (after acceptance by the publisher) become part of the original dataset.

4. Data-Flows at Global Publishers

Wolters Kluwer Germany (WKG) is an information service provider in the legal, business and tax domain. Business units of WKG are devided into "legal and regulatory" as well as as "tax and accounting" (see Fig.2). The business unit "legal and regulatory" serves mainly legal professionals in several different legal domains with content, software and services. WKG is headquartered in Cologne and has about 1,000 employees in 20 offices located across Germany. Wolters Kluwer Germany is part of Wolters Kluwer n.v., a global information services company with customers in the areas of legal, business, tax, accounting, finance, audit, risk, compliance and healthcare. In 2011, the company had annual revenues of 3.4 billion Euro and 19,000 employees worldwide with customers in over 150 countries across Europe, North America, Asia Pacific, and Latin America. Wolters Kluwer is headquartered in Alphen aan den Rijn, the Netherlands. Its shares are quoted on Euronext Amsterdam (WKL) and are included in the AEX and Euronext 100 indices.

Wolters Kluwer's strategy has 3 main focuses:

1. To deliver value at the point-of-use by helping customers to manage complex transactions to produce tangible results;
2. To expand solutions across whole processes, customers and networks;

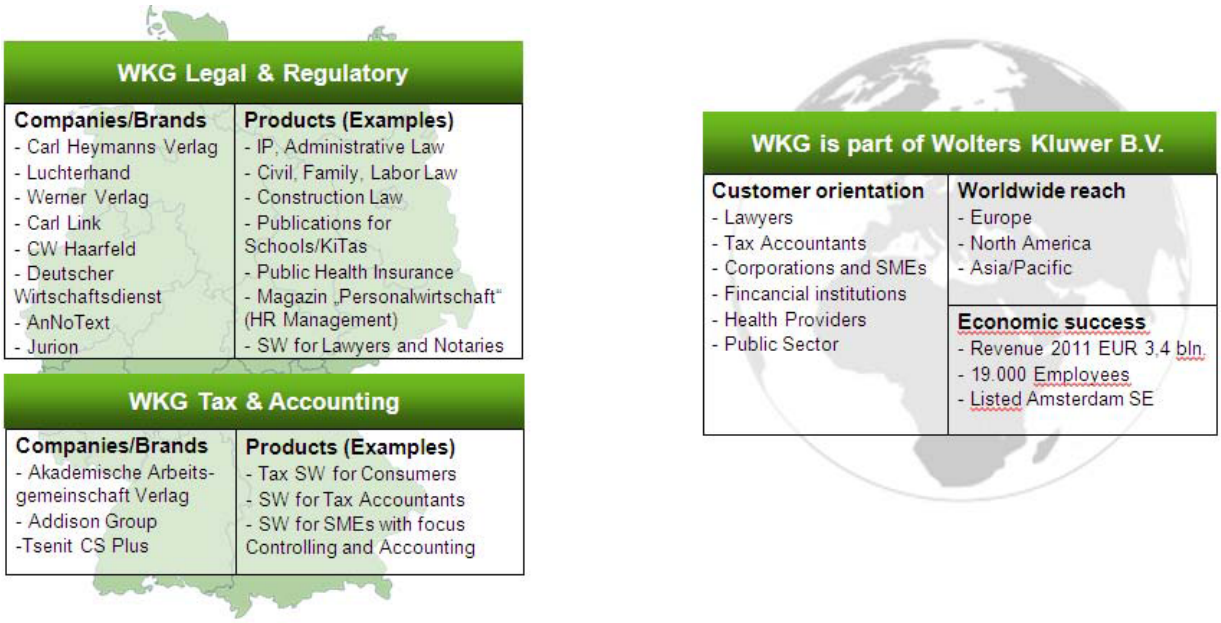

Fig. 2. Wolters Kluwer business units and products.

3. To raise innovation and effectiveness through global capabilities.

Assets like authoritative content, domain expertise and integrated workflow tools are basis of the strategic direction.

The application of semantic technologies at WKG was performed within the LOD2 EU project, which started in 2010. The WKG content supply chain in the beginning of the LOD2 project was quite typical for a publishing business (see Fig.3). It started with the process of content acquisition. Legal content was in general obtained from different external sources like public institutions (courts, ministries) or authors and then internally refined and consolidated by domain experts. The resulting authoritative content represented a valuable asset for specialized publishers like Wolters Kluwer but required, in its traditional form, a lot of resources.

Afterwards content was mostly manually classified, enriched and linked in further workflow steps by domain experts and technical writers. These actions could generate significant additional value for contents and their usage in different mediums and platforms. Subsequently, product managers collect contents for their products and compose or bundle them individually. But as the previous process of enhancing the contents was quite complex the selection and bundling was not really mature. Contents were therefore mainly used in one distinct product without additional informative value.

This fact had also impact on the sale process where products were barely connected with each other or with external open content. This limitation did also affect the sales process as well as the customer service.

5. Usage of the LOD2 Stack at WKD

When WKG, in particular the first authors of this application report, first investigated the paradigm of Linked Data, the respective lifecycle and the LOD2 stack supporting this lifecycle, we concluded the following: The Linked Data lifecycle is highly comparable to the existing workflows at Wolters Kluwer as an information provider; and the LOD2 stack can offer relevant functionality and technology complementary to our existing content management and production environments. Therefore the perception is that the impact of embracing the Linked Data paradigm and using Linked Data technology on our organization could be rather high.

During the course of testing Linked Data technologies, we learned more about the possibilities and current restrictions of the tool stack, but were also confronted with requirements from our internal business managers, which could hardly be solved within our existing technological infrastructure, but seemed to be

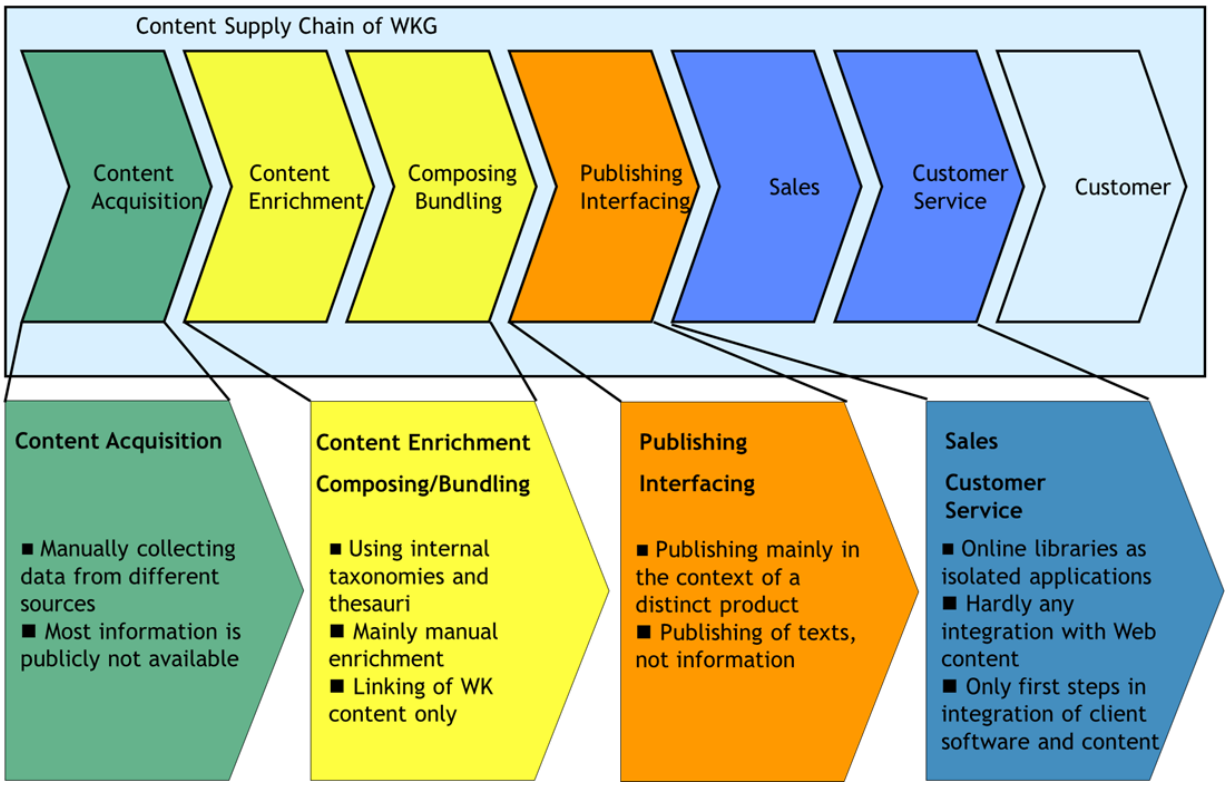

Fig. 3. Content supply chain at Wolters Kluwer.

perfectly suited for Semantic Web technology. Some of the major business requirements were:

1. Processing and enriching mass content from partner publishing houses into our products without having to handle this content within our standard content supply chain. This included the necessity to separate the source text from the (meta)data and the storage of the metadata in a dedicated repository. This also meant to rely as much as possible on unified controlled vocabularies, in order to ensure consistency across content sources.
2. Extension and consolidation of our controlled vocabularies. The extended usage of post search filters and typed auto-suggest functionalities required to work extensively on our controlled vocabularies. Once they were consolidated, we were able to offer better product features, but were also able to connect these vocabularies to external data sources and enrich our data even more.
3. Inclusion of product-specific variants of metadata associated to our documents: The transition from print to electronic contents led to the requirement that elements like "title" or some structural information needed to be different in the different media, even different in different electronic products - like on a desktop database vs. a mobile application. This should be handled independent from the textual sources, in order to ensure flexibility. Therefore, a central metadata repository was required.
4. Enabling a vertical view on content, based on a customer group specific angle (e.g. "law office" vs. "HR department in a company"). The diversification of electronic products implied the necessity to add quite different clusters of information connected to one piece of text, e.g. a law. This led even to the situation that the main complexity in the content processing was based on metadata handling and not on text handling anymore.

All these findings led in a rather early stage of the project to the conclusion that project results should immediately be evaluated by our internal technology and content experts in order to include these findings in our operational planning, so that we could ensure that

no resources were spent on legacy technology, where project results could be leveraged.

In general, there were two approaches for using tools from the LOD2 stack in an industrial environment: Taking the toolset and integrating it into our internal processes as open source software vs. approaching the vendor of the tool in order to license an enhanced commercial version of the tool. Since the technology used was quite new for us, we preferred the second option, which had the advantage that we were getting professional support and the assurance that maintenance and further development was guaranteed (e.g. *Virtuoso* and *PoolParty*). Of course, where no commercial tool was available for a given task, pure open-source was chosen, e.g. for linking we used the *SILK framework*.

One major task in the beginning of the project was to extract and transform our legal content. Mainly metadata, but also hyperlinks etc. were extracted from XML documents and knowledge bases capturing this were created. We used built-in functionality from Virtuoso and in addition the *Valiant* tool from the LOD2 Stack in order to execute XML to RDF conversion faster. This transformation task led to the finding that we have to differentiate between metadata that we want to represent using controlled vocabulary and metadata that is not explicitly controlled. This gave us the opportunity to introduce a concept for persistent IDs (in this case URIs) at WKG, which is an important basis for sustainable information management in general. The creation and usage of a controlled vocabulary on a broader basis also led to a consolidation and therefore enhanced the quality of our data, since a systematic curation process had to take place beforehand. We used PoolParty from the LOD2 stack for this curation process and the management and development of the controlled vocabulary. Since it is based on SKOS, it gave us all the freedom we needed for that purpose. For the rest of the data we used Virtuoso as a persistent storage facility, both for the original XML data as well as for the respective RDF.

Virtuoso and PoolParty were also the tools we chose for implementing a metadata database within our operational system, since the integration of both tools worked well and we could realize the main requirements for our application requirements:

We needed one single place to store and maintain rather different kinds of data. This data could come from the legal domain describing entities and knowledge, or it could come from internal (production) processes, covering things like process information or product variations.

We knew from our previous experience, that it was not possible to implement one stable and fixed relational data model, since the business requirements and the technical developments concerning data and metadata were changing so rapidly, that one main feature of the application needed to be "data model flexibility". Since there was no fixed schema necessary when using RDF, this requirement was easily met.

One major characteristic of the legal domain is the fact that legal matters and processes are highly connected to each other. This is, for example, reflected by a large number of explicit relations between documents. The preservation and usage of these relationships was of key importance, for example, to implement proper search capabilities. RDF gave us the possibility to easily establish these relationships.

The creation and maintenance of domain specific knowledge models required expertise and resources. In order to minimize this effort, one requirement was to be able to integrate external data sources in a controlled fashion as much as possible (e.g. DBpedia Live [10] information or publicly available domain thesauri). We used SILK and PoolParty for connecting our controlled vocabulary to external sources and the generic SILK framework for other metadata interlinking.

New usage scenarios for metadata were introduced, both from an internal process point of view as well as from a functionality point of view in our products. We needed to classify our documents according to legal domain structures. The knowledge represented in the metadata and their relations available in Virtuoso gave us the possibility to achieve the required classification quality. Another scenario was the qualified generation of an auto-suggest functionality, where the keywords shown to the user when typing his query were directly coming from our knowledge base and were therefore already normalized and prioritized.

The legal publishing market is still a national market. Currently, cross-border offerings are only relevant in certain areas like intellectual property law. However, this will change over time, especially within the European Union, where more and more legislation is performed on the European level. Having information available in RDF gives us the possibility to more easily align and connect to different Wolters Kluwer CMS systems in different countries in order to generate a comprehensive offering. This is already tested on a prototype level and will gain business importance

	Labour Law Thesaurus	**Courts Thesaurus**
Description	covers all main areas of labour law, like the roles of employee and employer; legal aspects around labur contracts and dismissal; also co-determination and industrial action	structures German and European courts in a hierarchical fashion and includes e.g. address information or map visualization
Concepts	1,728	1,499
Linked Sources	Standard Thesaurus für Wirtschaft, ZBW (zbw.eu/stw/), DBpedia, TheSoz from Leibniz Gesellschaft für Sozialwissenschaften (www.gesis.org), Eurovoc	DBpedia
Licenses	Data is licensed using 'Creative Commons Namensnennung 3.0 Deutschland (CC BY 3.0)' License, Data model is licensed using 'ODBL' License., Links to external sources are licensed using a 'CC0 1.0 Universal (CC0 1.0) Public Domain Dedication' License	
URL	`http://vocabulary.wolterskluwer.de/arbeitsrecht.html`	`http://vocabulary.wolterskluwer.de/court.html`
SPARQL endpoints	vocabulary.wolterskluwer.de/PoolParty/sparql/arbeitsrecht	vocabulary.wolterskluwer.de/PoolParty/sparql/court

Table 1

Description of the Labour Law and Courts thesauri published by WKG.

over time. One side effect of this development is that multilingual applications also gain importance. People search in their native language for content published in another language. Multilingual thesauri, which are already possible in the currently implemented infrastructure, will help us to address this issue.

The lack of public machine-readable legal resources in many European countries led to the decision to publish legal resources ourselves in order to initiate discussions within the publishing industry, but also within the linked data community and public bodies. These resources (cf. Table 1) are available via the SEMIC Semantic Interoperability Community platform[1] as semantic assets. This raised a lot of interest within the publishing industry, but also gave us the possibility to show our expertise to potential customers. Therefore this published data worked well also as a marketing tool.

We used the LOD2 Stack component OntoWiki as user interface for the presentation and maintenance of the metadata. It offers search and browse capabilities and is inherently based on RDF. For the operational environment, we chose a different strategy and were therefore not re-using LOD2 tools here. We had tools and interfaces in place for assigning metadata to content. In order to keep complexity away from the colleagues in the editorial departments, we decided to reuse and adapt these existing interfaces to the enhanced metadata assignment. On the other hand, the data in the metadata database gave us completely new possibilities for adding relations or analyzing data. This potential is by now not exploited yet and is on the agenda for one of the next development steps.

Tool support for maintaining knowledge models is a very prominent requirement. Therefore, in order to ensure and improve the quality of such models, the ontology enrichment and repair tool ORE is used by harmonising schema and instance data. The actual path for implementation and usage in an editorial domain for legal information needs to be evaluated still.

To sum up, the objectives to deploy tools from the LOD2 stack in our operational systems were mainly targeting at making our internal content processes more flexible and efficient, but also targeted new feature for WKG's electronic and software products. Once the technological basis was laid, immediately new opportunities for further enhancements showed up, so that this new part of our technical infrastructure already gained importance and there is no doubt, that this process will continue. The tools currently used from the LOD2 stack are well integrated with each other, which enables an efficient workflow and processing of information. URIs in PoolParty based on controlled vocabularies are used by Valiant for the content transformation process and stored in Virtuoso, so that it can easily be queried via SPARQL and displayed in OntoWiki. The usage of the LOD2 stack as such has the major advantage that the installation is easy and the issues around different versions not working smoothly

[1] http://semic.eu

with each other are avoided. All this are major advantages compared to the separate implementation of individual tools.

A major challenge, however, is not the new technology as such, but a smooth integration of this new paradigm in our existing infrastructure and a stepwise replacement of old processes with the new and enhanced ones.

To summarise the application from a software perspective, we list LOD2 Stack components which were deployed in the content production and management processes at WKG (cf. Figure 4):

1. *Valiant* is an extraction/transformation tool that uses *XSLT* to transform XML documents into RDF. The tool can access data from the file system or a WebDAV repository. It outputs the resulting RDF to disk, WebDAV or directly to an RDF store. For each input document a new graph is created.
2. *Virtuoso* [4] is an enterprise grade multi-model data server. It delivers a platform agnostic solution for data management, access, and integration. Virtuoso provides a fast quad store with SPARQL endpoint and WebID support.
3. *PoolParty* [11] is a tool to create and maintain multilingual *SKOS* (Simple Knowledge Organisation System) thesauri, aiming to be easy to use for people without a Semantic Web background or special technical skills. PoolParty is written in Java and uses the SAIL API, whereby it can be utilized with various triple stores. Thesaurus management itself (viewing, creating and editing SKOS concepts and their relationships) can be performed in an *AJAX* front-end based on the Yahoo User Interface (YUI) library.
4. *OntoWiki* [2] is a PHP5 / Zend-based Semantic Web application for collaborative knowledge base editing. It facilitates the visual presentation of a knowledge base as an information map, with different views of instance data. It enables intuitive authoring of semantic content, with an inline editing mode for editing RDF content, similar to WYSIWYG for text documents.
5. *Silk* [5] is a link discovery framework that supports data publishers in setting explicit links between two datasets. Using the declarative Silk - Link Specification Language (Silk-LSL), developers can specify which types of RDF links should be discovered between data sources as well as which conditions data items must fulfill in order to be interlinked. These link conditions may combine various similarity metrics and can take the graph around a data item into account using an RDF path language.
6. *ORE* [8] (Ontology Repair and Enrichment) allows knowledge engineers to improve an OWL ontology or SPARQL endpoint backed knowledge base by fixing logical errors and making suggestions for adding further axioms to it. ORE uses state-of-the-art methods to detect errors and highlight the most likely sources for the problems. To harmonise schema and data in the knowledge base, algorithms of the DL-Learner [6,7,9] framework are integrated.

An important asset of the LOD2 stack is the fact, that components do interact and support specific steps of the data transformation lifecycle and management process. Figure 4 shows the already operationally implemented interplay of LOD2 stack components in the processes of WKD.

A document, e.g. a law, is transformed from XML into RDF with the Valiant Tool. The RDF (meta)data is stored in Virtuoso and managed using a customized version of Ontowiki. Examples for such data in legislations are the "legislation date", "law abbreviation" or "legislation type". Technical editors / domain experts can within this environment add further metadata or change existing ones via editing features.

Controlled vocabularies are managed in Poolparty. Thesaurus concepts can be created, ordered and edited in the system. Both metadata management systems interact in a way that e.g. vocabularies from Poolparty can serve filtering functionalities for documents in OntoWiki. This way, we can browse in the navigation pane for a specific "legislation type", an "area of law", "authors" or "courts".

External sources are linked using the Silk framework to document metadata in OntoWiki on the one hand and with controlled vocabularies in Poolparty on the other hand. In case of a law, there could be an enrichment of the document by the area of law or the jurisdiction (i.e. geographical coverage) of a law. Vocabularies can be enriched by abstracts or synonyms. This linking enriches the documents and supports further functionality, especially in the case of controlled vocabularies.

For quality control ORE and the included DL-Learner library is used. In a first step, ORE analyzes the existing instance data and suggests class descriptions for each class contained in the domain ontology.

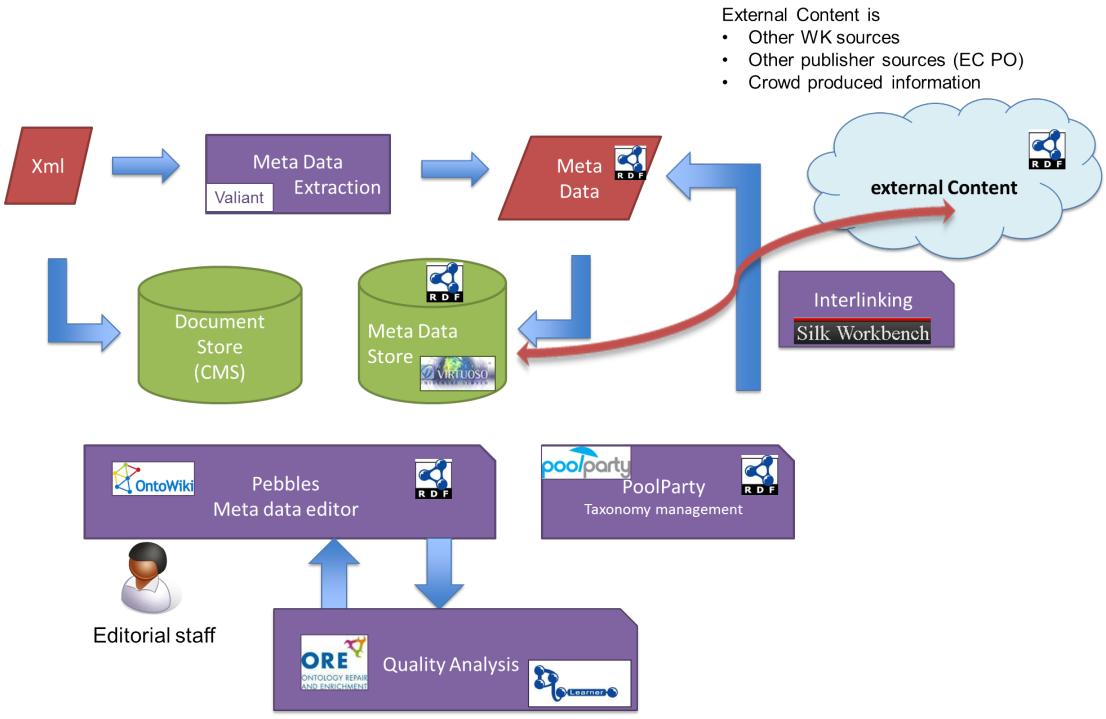

Fig. 4. Content management process at WKG and usage of LOD2 stack components.

For instance it suggested that each document of type "aufsatz" (German) has at least a title, an editor or creator, as well as information about the start and end position in the page. Based on the suggestions, a domain expert creates and refines the schema. Afterwards, the axioms in the schema are used as constraints and converted into SPARQL queries, which allows for the detection of instance data that does not fullfil the requirements, e.g. finding instances of "aufsatz" where the title is missing. Technically, expressive OWL schema axioms are used as constraints by employing a closed world assumption via a closed world assumption.

6. Related Work

There are other publishers that already use semantic technologies to enhance their own online publishing processes, although we believe that WKG applies a richer set of tools covering many Linked Data lifecycle stages. The *New York Times*, one of the largest American daily newspapers, publishes its index as RDF since 2009. About 10,000 concepts as persons, organizations, locations or descriptors that are used for tagging articles are published under a CC-BY license with respective metadata and links to DBpedia, to Freebase or into the Times API[2]. The BBC, a British public service broadcasting statutory corporation uses their *dynamic semantic publishing architecture* to enhance content processing workflows for their website. Contents are embedded in an ontological domain-modelled information architecture that enables automated aggregation processes and publishing as well as re-purposing of interrelated content objects.[3]

7. Future Work

Ingestion of more data in general and the inclusion of more external information is one major topic, but also preparing this conglomerate of information for real world usage in an industrial environment is a major challenge. The latter covers e.g. issues around quality, governance and licensing. In detail, we will focus on the following tasks:

[2]See `http://data.nytimes.com/`.
[3]See `http://www.bbc.co.uk/blogs/bbcinternet/2012/04/sports_dynamic_semantic.html`.

First of all, we will work on the further deployment and adoption of the LOD2 tool stack to further enhance our metadata sets. We will focus on repair and enhancements of the existing RDF schema, automatic classification of contents, entity extraction and improved functionality of the metadata management tools. We are especially aiming for improvements of metadata management workflows by enhanced usability or functionality in order to fasten semi-automatic processes.

Concerning the interlinking processes, we want to explore new external sources such as *GND* (`http://datahub.io/dataset/dnb-gemeinsame-normdatei`) for enrichment of our datasets and investigate new opportunities to improve the generation and quality of new vocabularies.

In order to improve the interface and authoring options, we will invest in enhancements in publishing, search, browsing and exploring of metadata sets within metadata management.

The topic of licensing strategies, practices and recommendations is another important area, in particular when datasets from various sources licensed under different licensing schemes are fused.

Acknowledgements

This work was supported by grants from the European Union's 7th Framework Programme provided for the projects LOD2 (GA no. 257943) and GeoKnow (GA no. 318159).

References

[1] S. Auer, L. Bühmann, C. Dirschl, O. Erling, M. Hausenblas, R. Isele, J. Lehmann, M. Martin, P. N. Mendes, B. van Nuffelen, C. Stadler, S. Tramp, and H. Williams. Managing the lifecycle of linked data with the lod2 stack. In *Proceedings of International Semantic Web Conference (ISWC 2012)*, 2012. 22

[2] S. Auer, S. Dietzold, and T. Riechert. OntoWiki - A Tool for Social, Semantic Collaboration. In *5th Int. Semantic Web Conference, ISWC 2006*, volume 4273 of *LNCS*, pages 736–749. Springer, 2006.

[3] S. Auer and J. Lehmann. Making the web a data washing machine - creating knowledge out of interlinked data. *Semantic Web Journal*, 2010.

[4] O. Erling. Virtuoso, a hybrid rdbms/graph column store. *IEEE Data Eng. Bull.*, 35(1):3–8, 2012.

[5] A. Jentzsch, R. Isele, and C. Bizer. Silk - generating rdf links while publishing or consuming linked data. In *ISWC 2010 Posters & Demo Track*, volume 658. CEUR-WS.org, 2010.

[6] J. Lehmann. DL-Learner: learning concepts in description logics. *Journal of Machine Learning Research (JMLR)*, 10:2639–2642, 2009.

[7] J. Lehmann, S. Auer, L. Bühmann, and S. Tramp. Class expression learning for ontology engineering. *Journal of Web Semantics*, 9:71 – 81, 2011.

[8] J. Lehmann and L. Bühmann. Ore - a tool for repairing and enriching knowledge bases. In *9th Int. Semantic Web Conference (ISWC2010)*, LNCS. Springer, 2010.

[9] J. Lehmann and P. Hitzler. Concept learning in description logics using refinement operators. *Machine Learning journal*, 78(1-2):203–250, 2010.

[10] M. Morsey, J. Lehmann, S. Auer, C. Stadler, and S. Hellmann. DBpedia and the Live Extraction of Structured Data from Wikipedia. *Program: electronic library and information systems*, 46:27, 2012.

[11] T. Schandl and A. Blumauer. Poolparty: Skos thesaurus management utilizing linked data. In *7th Extended Semantic Web Conf., ESWC 2010*, volume 6089 of *LNCS*, pages 421–425. Springer, 2010.

[12] S. Tramp, P. Frischmuth, T. Ermilov, and S. Auer. Weaving a Social Data Web with Semantic Pingback. In P. Cimiano and H. Pinto, editors, *Proceedings of the EKAW 2010 - Knowledge Engineering and Knowledge Management by the Masses; 11th October-15th October 2010 - Lisbon, Portugal*, volume 6317 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 135–149, Berlin / Heidelberg, October 2010. Springer.