Semantic Web 1 (2016) 1–5
IOS Press


# Facilitating Data-Flows at a Global Publisher using the Linked Data 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], Bert Van Nuffelen [d]

[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*
[d] *TenForce, Havenkant 38, 3000 Leuven, Belgium*
*E-mail: bert.van.nuffelen@tenforce.com*

**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 Linked Data 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 Linked Data Stack at a global publishing company. We give an overview over the Linked Data 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, Linked Data 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. For instance, 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) and therefore the internal workflow processes especially for those publishers targeting professional audiences (e.g. legal, tax, accounting professionals) have to be adapted in order to provide this high value content.

Let's motivate our work in more detail with a 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

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

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 (enterprise resource planning) 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 have 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 Linked Data and tools from the Linked Data Stack can address those challenges. This application report builds on [1]. While the previous article focused on a detailed characterisation of the Linked Data Stack, this application report focuses on a detailed description of the usage scenarios at a global publisher.

In the past publishers were the driving force behind document representation formats as SGML and XML. Using that technology usually a document centric content manament system is established. Today this technology is still a corner stone of many publishers however the aforementioned rise in electronic publishing and electronic usage of documents challenges (the document centric usage of) this technology. Electronic publishing requires separation of the core content (represented as XML) and metadata, smaller partitioning of the content and the ability to included life external information.

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 and machine learning techniques to enrich the knowledge bases with ontological structures. Since the original documents might change (like a new version of a law), we need processes for maintaining and further developing the extracted knowledge. Finally, semantic search, exploration and visualization techniques can help to gain new insights from the semantically represented content. The Linked Data 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.

The application report is structured as follows: In Section 2, we present the Linked Data 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 data-flows at global publishers, which are described in Section 4. Based on this, in Section 5, we describe how the Linked Data 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 Linked Data Stack

The description of the Linked Data Stack (formaly known as the LOD2 stack) and the Linked Data lifecycle (see Figure 1) are extensions of earlier work in [1] and [3]. The Linked Data 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 Linked Data stack is available at `http://stack.linkeddata.org` The major components

of the Linked Data Stack are open-source facilitating a wide deployment potential. Through an iterative 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: by exploiting the Linked Data (RDF, SPARQL, OWL) paradigm as the main application interface the plugging in of alternative (third-party) implementations is enabled.

In order to fulfill these requirements, the architecture of the Linked Data 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 Linked Data Stack components. Using the Debian system also allows to facilitate the deployment of the Linked Data Stack on individual servers, cloud or virtualization infrastructures.
2. The use of RDF as the data representation format and SPARQL as the data exchange mechanism creates a uniform and universal knowledge exchange bus. In its simplest form the bus is a central local RDF store, however due to the built-in distribution nature of RDF and SPARQL more complex setups are easily realized. All components of the Linked Data Stack access via the bus the knowlegde 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 exploit vocabularies. Vocabularies (or ontologies) provide semantics to the data for the information domain they cover. Using vocabularies allows information exchange beyond raw data.
3. Integration of the Linked Data Stack user interfaces based on REST API's. The available components form a collection of user interfaces that are technologically and methodologically quite heterogeneous. The Linked Data Stack does not aim to resolve this heterogeneity, since each tool's UI is specifically tailored for a certain purpose. Instead, we develop a common entry points for accessing and managing the data via REST API's that offer dedicated support for one task: e.g. selecting a graph, or validating a dataset, etc. This approach fosters both the exploration of new ideas and new functionalities as the reuse of already developed functionalities in other applications.

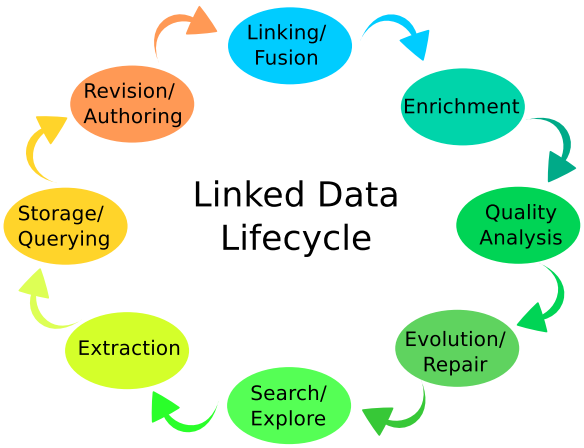


Fig. 1. Stages of the Linked Data life-cycle supported by the Linked Data Stack.

These three pillars comprise the methodological and technological framework for integrating the very heterogeneous Linked Data 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:

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.
2. *Authoring:* The Linked Data 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.
3. *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.

4. *Enrichment:* 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 Linked Data Stack, semi-automatic components for this purpose are included.
5. *Quality:* The quality of content on the Data Web varies, as the quality of content on the document web varies. The Linked Data Stack comprises techniques 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.
6. *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 Linked Data Stack comprises methods to spot problems in knowledge bases and to automatically suggest repair strategies.
7. *Search/Browsing/Exploration:* For many users, the Data Web is still invisible below the surface. Therefore search, browsing, exploration and visualization techniques for different kinds of Linked Data (i.e. spatial, temporal, statistical) are developed making the Data Web sensible for real users.

We refer to [11] for detailed explanations and specific examples for each of those stages. These lifecycle 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 (synergies) 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, the Linked Data Stack results in 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).

Linked Data is as technological foundation a firm basis to incrementally grow the interaction pattern. One can start with the minimal data flow sufficient to capture the core data management problem. Later on the data flow can be extended with new interaction points creating a higher value data chain. The easy extendibility of data flows allows to incorporate recent developed tools and techniques so that at any time an data flow is served by the best component. For instance, a typical first step of Publishers into Linked Data starts with the representation of their controlled vocabularies as SKOS vocabularies and updating their content using the concept URI's instead of internal identifiers. In a second step the data flow can be extended with automated annotation of the content using those SKOS vocabularies and so on. This approach allows to tackle the data value chain in a non-disruptive way, enhancing it on a by-need basis.

## 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;
3. To raise innovation and effectiveness through global capabilities.

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

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

Afterwards content is mostly manually classified, enriched and linked in further workflow steps by domain experts and technical writers. These actions generate significant additional value for contents and their usage in different mediums and platforms. Subsequently, product managers collect content for their products and compose or bundle them individually. But as the previous process of enhancing the contents is quite complex and labour intensive, the selection and bundling is often limited to the most obvious and necessary actions. Contents is therefore mainly provided via one distinct product without additional informative value to the customers. This fact impacts the sale process where only products are offered that are barely connected with each other or offered with external content. So in order to provide a better customer service, the core content supply chain has to be improved.

## 5. Usage of the Linked Data Stack at WKG

When WKG, in particular the first authors of this application report, first investigated the paradigm of Linked Data, the respective lifecycle and the Linked Data 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 Linked Data stack offers relevant functionality and technology complementary to the existing content management and production environments.

It was decided to explore the impact of Linked Data by setting up pilots using the Linked Data stack. Each of the pilots tackled a crucial functionality or new opportunity. Using this approach experience, usage and developement guidelines, best practices and current restrictions of the components available in the Linked Data stack were collected. The pilots not only challenged the Linked Data stack, but also the WKG business managers. They provided requirements for challenges that could hardly be solved with the existing technological infrastructure. 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 the 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 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

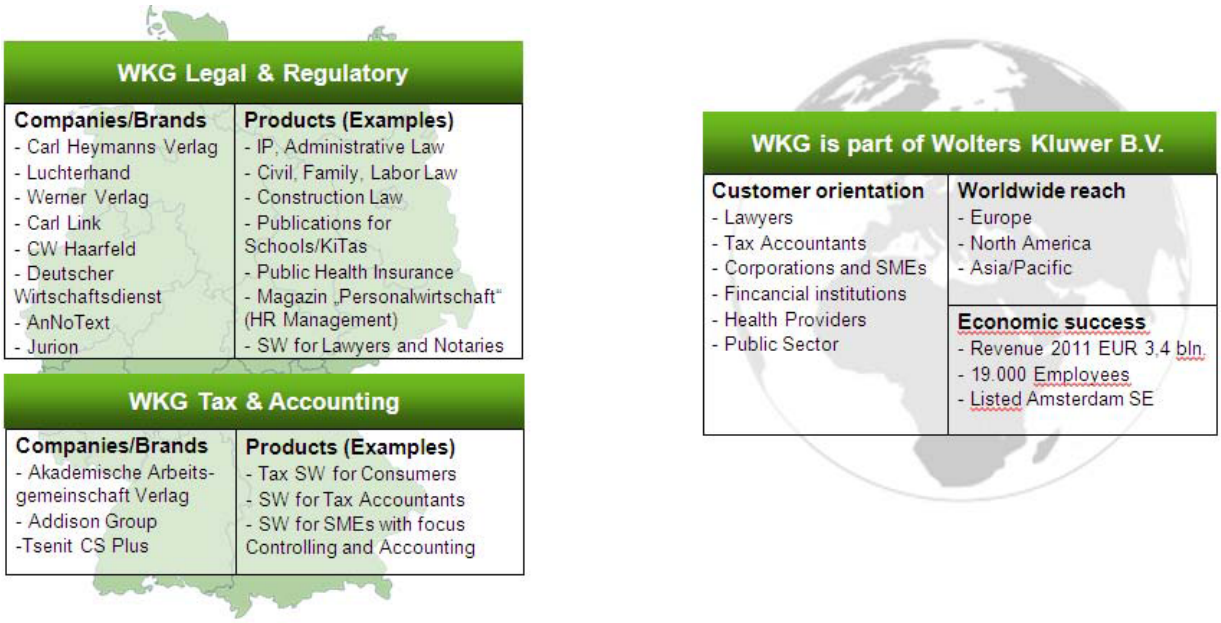


Fig. 2. Wolters Kluwer business units and products.

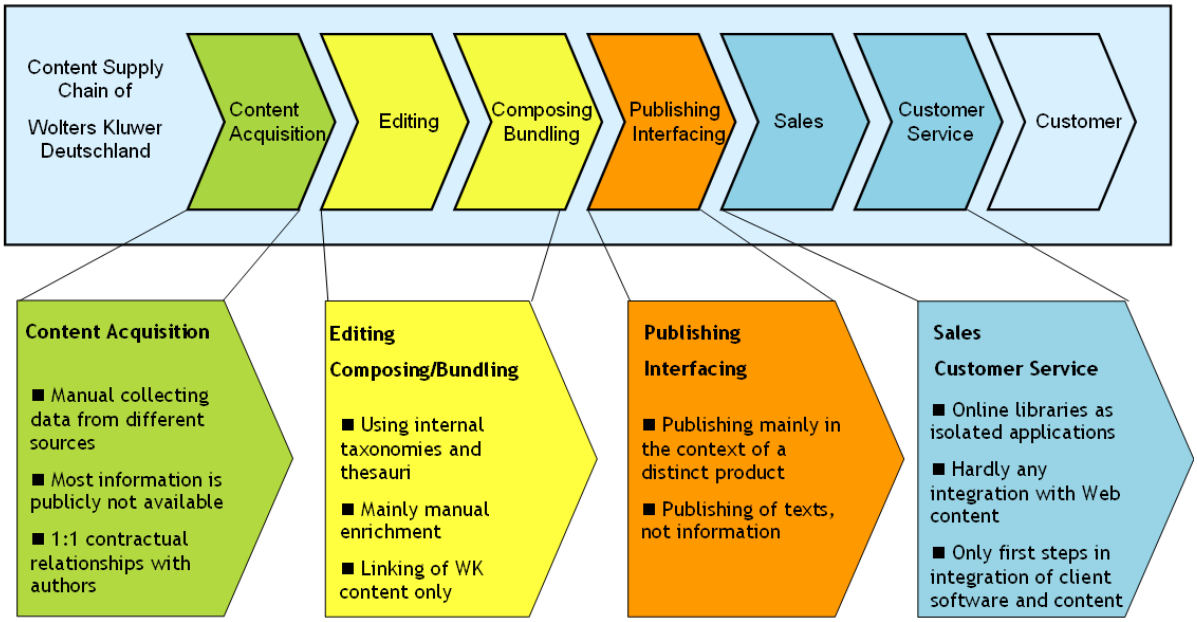


Fig. 3. Content supply chain at Wolters Kluwer.

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.

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 diversi-

> fication 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.

Already in an early stage of the LOD2 project it was observed that project results should immediately be evaluated by the WKG internal technology and content experts. By taking up the findings and leverage the project results in the operational planning, the resource spending in lecacy technology could be reduced.

In general, there were two approaches for using tools from the Linked Data 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. Since WKG did not have any internal know-how on this technology, WKG preferred to partner with commercial vendors when available, 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*). In all other cases pure open-source was chosen, e.g. for linking we used the *SILK framework*.

Like each data management project, the work in the beginning was to extract from the provided raw legal content the required information into RDF format. Each provided XML document was converted into a knowledge base containing the metadata (e.g. title, author, reference number, ...) and structural information (e.g. chapter, sections, paragraphs, ...). The actual legal clause texts were initially not included, however for a later excercise the extraction process has been updated to include them too.

For quality assurance en focussing reasons, it was decided to iterative refine the extraction rules. The process of transformation rules writing, processing documents, reviewing and step-wise extensions is labor intensive work that must be supported with adaquate tooling. In our case, the provided amount of legal documents was sizable. Because of the document oriented conversion approach, the scaling has been built around this dimension. The tool chain consists of the *Valiant* tool parallelizing the processing of XML documents to RDF, in combination with the bulk loading of Virtuoso.

A lesson learned from this transformation task was that we have to differentiate between metadata that we wanted to represent using controlled vocabularies and metadata that is not explicitly controlled. In case there is a controlled vocabulary, the data extraction chain has to be extended with a reconsiliation step. The provided information sometimes only covers the label instead of the identifier of the concept. This situation has been dealt with by applying a reconsilation step in which the label is replaced with a proper reference to the targetted identifier. In case the reconsilation service did provided none or multiple answers the orginal extracted data was kept in the resulting data. By quering the knowledge base afterwarts these cases were retrieved and used to improve the controlled vocabularies or to correct the legal documents. This investigation process naturally drives to higher quality data. The management of the controlled vocabularies and the associated reconsiliation service are provided by the *PoolParty* solution.

The assessment of candidates for controlled vocabularies revealed that although they look natural candidates the creation of a taxonomy; the actual needed effort to create a correct disambiguated list was beyond the project budget. For instance the author list showed that there are several authors with the same name and that they have to be disambiguated per document.

The convertion into properly managed taxonomies in PoolParty for those lists that were selected, raised the fundamental Linked Data concept of persistent identifiers to be considered by WKG. Without a URI strategy in place sustainable information management is impossible.

The outcome of the extraction process: the original XML documents and the associated extracted metadata as RDF are published using Virtuoso on a SPARQL endpoint. The PoolParty publishing features have been used to publish the controlled vocabularies as SKOS on another SPARQL endpoint.

Based on the extraction work, WKG decided to integrated Virtuoso and PoolParty a metadata management solution within our operational system. This metadata management solution realizes the main requirement for further exploitation of the WKG legal content beyond the current practice: namely a central single storage of metadata. This central data storage does not include only the metadata already available in the legal documents but also supporting information from internal and external sources. For instance, product variations or publication status information can be stored.

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. DBpedia data were meaningful and helpful for concrete use cases. However, some of the data (e.g. geolocations) have been transformed incorrectly to DBpedia and caused errors when using it. These issues were communicated to the community and got resolved.

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, there indications it will change over time, especially within the European Union, where more and more legislation is performed on the European level. Having information available as machine readible Linked Data gives publishers the possibility to more easily align and interconnect different CMS systems used by the enterprise entities in different countries in order to generate a comprehensive offering. This is already tested on a prototype level and will gain business importance 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 part of the above discussed infrastructure, play a central role to address this issue.

Despite the growing amount of available Linked Data, the absence of sufficient (public) machine-readable legal resources in many European countries led to the decision by WKG to publish legal resources. It served as an excellent marketing tool showing WKG's expertise, but more importantly this initiated discussions within the publishing industry, but also within the linked data community and public bodies. An outcome of this increased interest is that this work has been used to movitate the adoption of the SKOS standard as a recommended standard (comply or explain standard) for publishing taxonomies by the Dutch government.

We used the Linked Data 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. In contrast to other components introducing this frontend in the WKG operational environment is more complicated because it affects the way of working throughout the entire company especially for the content editors. Therefore a different strategy has been chosen for transfering the experience to the existing editorial environment. It has been decided to re-use and adapt the existing interfaces to include the enhanced metadata assignment. This approach limits the impact on the editorial departments. For completely new tasks, outside the existing work environment of the editorial staff, like adding relations (e.g. courtnames with geodata) or analyzing data (like popular legal topics in court cases over time), direct access to the metadata database is our preferred approach.

Tool support for maintaining knowledge models is a very prominent requirement. Within the Linked Data stack the ontology enrichment and repair tool ORE provides support in this area. Our experiment setup objective was to harmonize the schema and the instance data. Although some situations could be tackled, the experiment was to limited to determine how this kind of tool can be used in the domain of legal information. Further investigation is required.

To sum up, the objectives to deploy tools from the Linked Data Stack in WKG's operational systems were mainly targeting at making our internal con-

| | **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.

tent 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 Linked Data 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 Linked Data Stack as such has the major advantage that the installation is easy and the issues around different versions not working smoothly 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 Linked Data Stack components which were deployed in the content production and management processes at WKG (cf. Figure 4; general statistics in Table 3):

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* [12] 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.

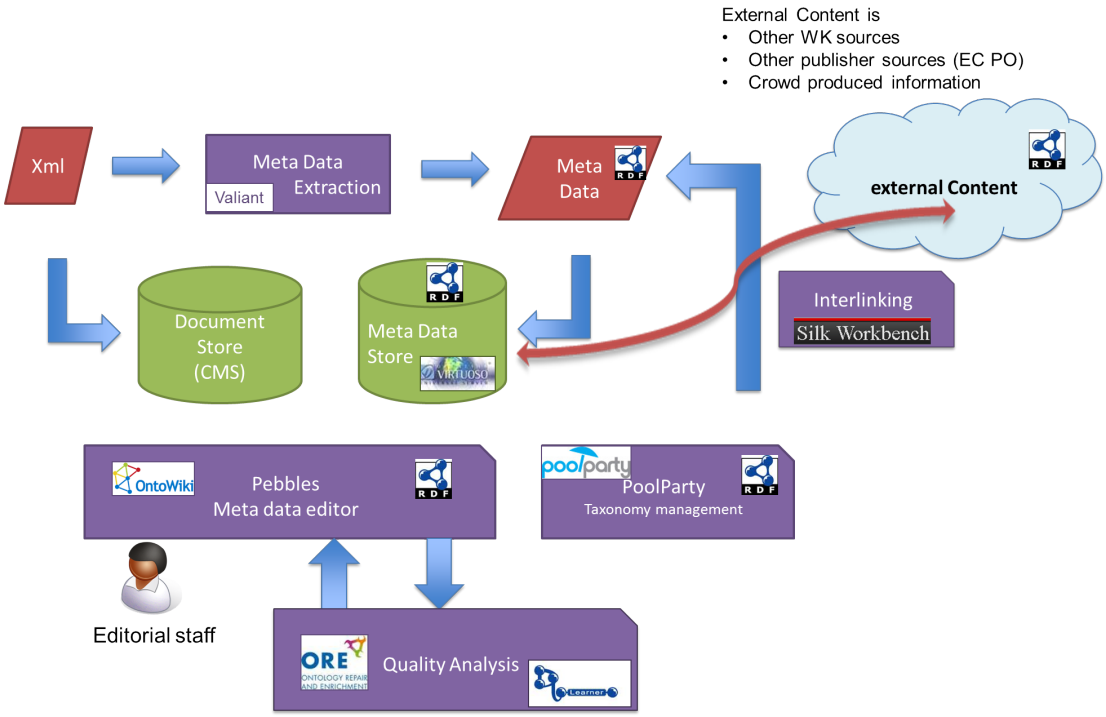


Fig. 4. Content management process at WKG and usage of Linked Data Stack components.

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 Linked Data 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 Linked Data Stack components in the processes of WKG.

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 (for document statistics see Table 2) . 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 (for external sources see Table 4). In case of a law, there could be an enrichment of the doc-

Table 2
Frequently used document classes.

| Top level document type URI | Document Count |
|---|---|
| all documents | 674884 |
| http://schema.wolterskluwer.de/aufsaetze | 218186 |
| http://schema.wolterskluwer.de/beitraege | 24611 |
| http://schema.wolterskluwer.de/entscheidungen | 398398 |
| http://schema.wolterskluwer.de/kommentare | 26653 |
| http://schema.wolterskluwer.de/lexikon | 47 |
| http://schema.wolterskluwer.de/rezensionen | 99 |
| http://schema.wolterskluwer.de/vorschrift | 6890 |

Table 3
Overall statistics.

| Measure | Count |
|---|---|
| Triples | 42969471 |
| Graphs | 830090 |
| Entities | 6812303 |

ument 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. For instance it suggested that each document of type "aufsatz" (German expression for article) 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[1]. 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.[2]

Table 4
Link statistics.

| Link Target | Count |
|---|---|
| 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 |
| WKG Labor Law Thesaurus to WKG subjects | 70 |

## 7. Future Work

Ingestion of more data in general and the inclusion of more external information especially 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:

First of all, we will work on the further deployment and adoption of the Linked Data 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.

[1] See http://data.nytimes.com/.

[2] See http://www.bbc.co.uk/blogs/bbcinternet/2012/04/sports_dynamic_semantic.html.

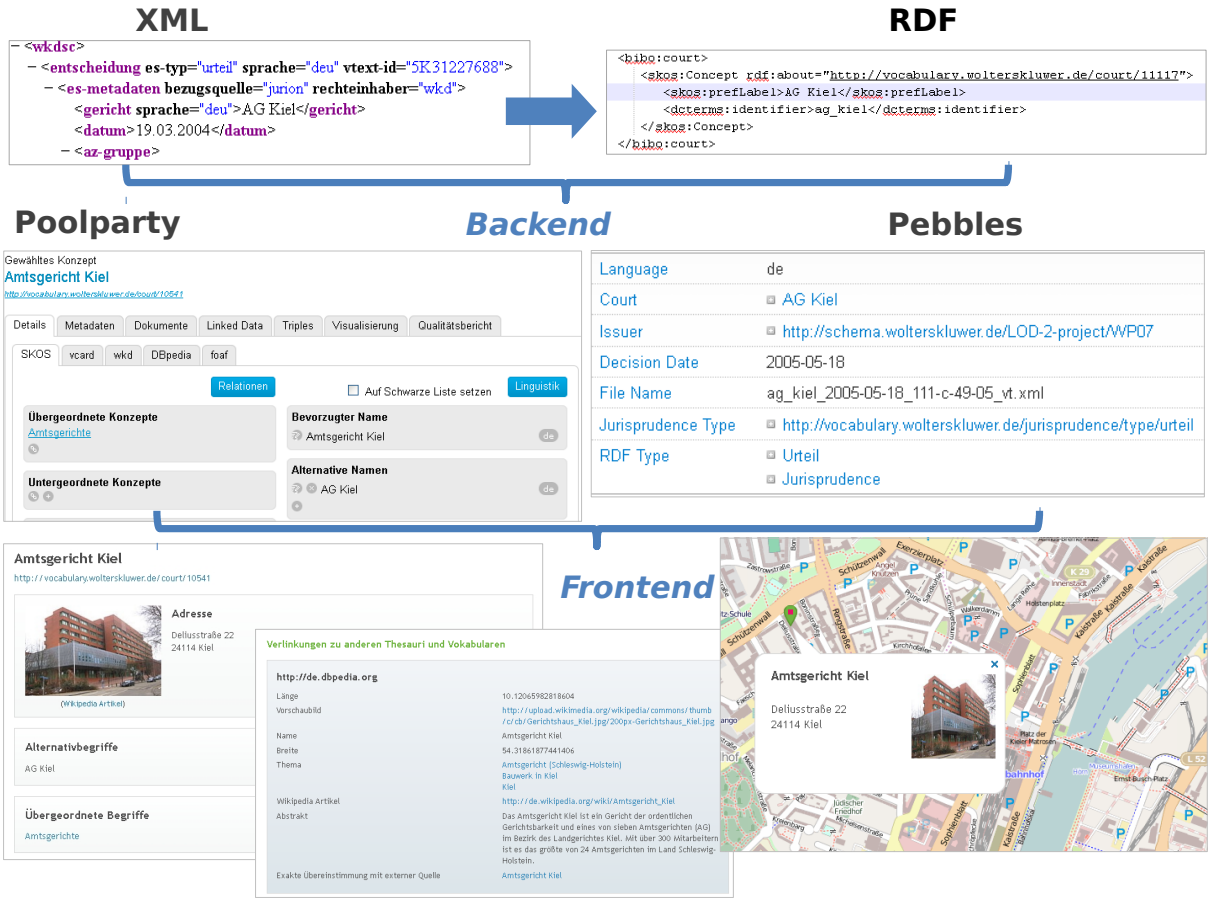


Fig. 5. Workflow illustrating how data was transformed from XML to RDF, enriched via Poolparty and Pebbles and then published in various user interfaces.

Concerning the interlinking processes, we plan exploring 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 (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] A.-C. N. Ngomo, S. Auer, J. Lehmann, and A. J. Zaveri. Introduction to linked data and its lifecycle on the web. In *Reasoning Web*. 2014.

[12] 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.