Supporting the Data Lifecycle at a Global Publisher using the Linked Data Stack

by Christian Dirschl, Katja Eck and Jens Lehmann

The Linked Data Stack is an integrated distribution of aligned tools that support the whole lifecycle of Linked Data from extraction, authoring/creation via enrichment, interlinking and fusing through to maintenance. A global publishing company represents an ideal recent real-world usage scenario, illustrating the Linked Data Stack and the underlying lifecycle of Linked Data (including data-flows and usage scenarios).

In these times of omnipresent electronic devices, the ways of consuming information are changing and so too are the expectations of customers. Documentation and processing of publishers' data, however, are lagging behind.

Let's assume an accounting professional is working for a leading consultancy and is responsible for certifying tax returns for an international customer. In the future, the publisher aims to deliver more personalized and context specific information precisely fulfilling his need to track changes in information provided by a multitude of sources.

The Linked Data Stack provides specialized tools for each Linked Data lifecycle stage (e.g. data enrichment, management of knowledge bases, reasoning techniques and semantic search support) and can consequently be used to facilitate the semantic content processing workflows.

The Linked Data Stack and Life-Cycle

The description of the Linked Data stack and the Linked Data lifecycle are based on earlier work in [1] and [2]. The Linked Data Stack is an integrated distribution of aligned tools that support the whole lifecycle of Linked Data from extraction, authoring/creation via enrichment, interlinking and fusing through to maintenance. The major components of the Linked Data Stack are open-source in order to facilitate wide deployment. The stack is designed to be versatile; for all functionalities there are clear interfaces, which enable the plugging in of alternative third-party implementations.

In order to fulfill these requirements, the architecture of the Linked Data Stack is based on the following basic principles:
• 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 facilitates the deployment of the Linked Data Stack on individual servers, cloud or virtualization infrastructures.
• Use of a central SPARQL endpoint and standardized vocabularies for knowledge base access and integra-

Figure 1: Overview of the stages of the Linked Data lifecycle [3].

tion between different tools: All components of the Linked Data 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 lifecycle.

Usage of the Linked Data Stack at WKD

Wolters Kluwer Germany (WKD) is an information service provider in the legal, business and tax domain. WKD is part of global Wolters Kluwer n.v. In 2012, the company had an annual revenue of 3.6 billion Euro,19,000 employees worldwide and customers in over 150 countries across Europe, North America, Asia Pacific and Latin America.

The paradigm of Linked Data and its lifecycle is highly compatible with the existing workflows at Wolters Kluwer as an information provi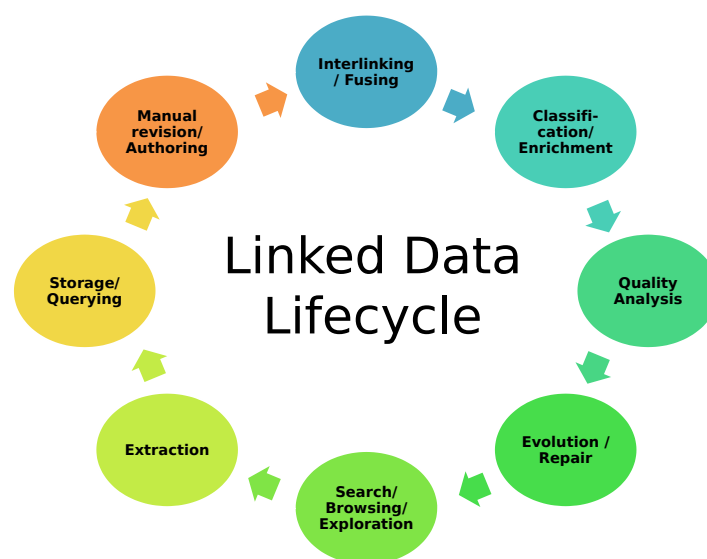der; consequently the Linked Data stack can offer functionality and technology that is relevant and complementary to the existing content management and production environment.

The main aim of implementing tools from the Linked Data stack into WKD's operational systems was to make the internal content processes more flexible and efficient, but feature requirements of the company's electronic and software products also had to be taken into consideration. Once the technological basis was laid, opportunities for further

Figure 2: Publishing workflow of Wolters Kluwer and Linked Data stack components [3]
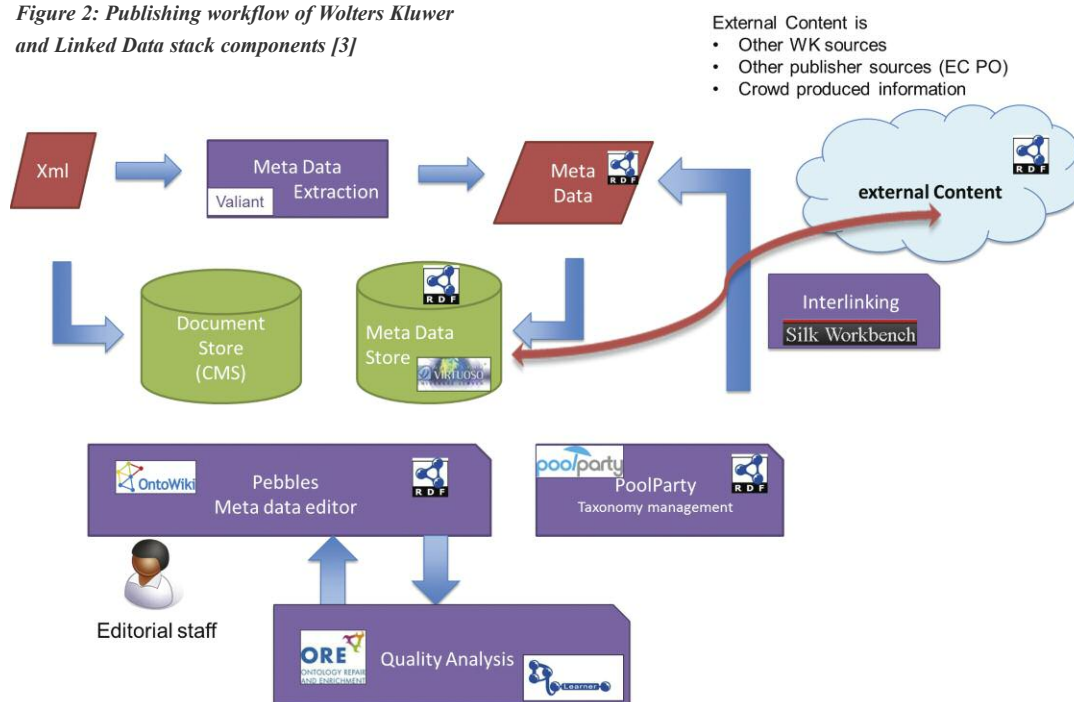
enhancements were immediately revealed; thus the Linked Data stack proved its value from early on, and there is no doubt that its importance will only continue to grow. The tools currently used from the Linked Data stack are well integrated with one other, which enables an efficient workflow and processing of information. URIs in PoolParty based on controlled vocabulary are used by Valiant for the content transformation process, and stored in Virtuoso, for easy querying via SPARQL and display in OntoWiki.

Using a Linked Data stack has the major advantage that installation is easy and issues associated with different versions not working smoothly together are avoided. These represent major advantages compared to the separate implementation of individual tools. Figure 2 shows the interplay of partially operational Linked Data stack components in the processes of WKD.

The major challenge, however, is not the new technology per se, but a smooth integration of this new paradigm into WKD's existing infrastructure and a stepwise replacement of old processes with the new and enhanced ones.

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 are available via the SEMIC Semantic Interoperability Community platform as semantic assets, but also directly at vocabulary.wolterskluwer.de.

Future Work
In the future, we will concentrate on adding further tools from the tool stack into our internal content processing engine as well as adding additional external sources to our knowledge base. These steps will be accompanied by detailed user and usability tests as well as documentation of business impact.

Links:
Linked Data Stack Website:
http://stack.linkeddata.org
Wolters Kluwer Germany:
http://www.wolterskluwer.de
LOD2 project: http://lod2.eu
GeoKnow project: http://geoknow.eu
Thesauri:
http://vocabulary.wolterskluwer.de/court.html
http://vocabulary.wolterskluwer.de/arbeitsrecht.html

References:
[1] S. Auer et al.: "Managing the life-cycle of Linked Data with the LOD2 Stack" in proc. of the 11th International Semantic Web Conference (ISWC), Springer, 2012, dx.doi.org/10.1007/978-3-642-35173-0_1
[2] S. Auer, J. Lehmann: "Making the Web a Data Washing Machine - Creating Knowledge out of Interlinked Data", Semantic Web Journal, 1,1-2, pp 97-104, IOS Press, 2010, http://www.semantic-web-journal.net/sites/default/files/swj24_0.pdf
[3] C. Dirschl et al.: "Facilitating Data-Flows at a Global Publisher using the LOD2 Stack"; submitted to the Semantic Web journal, http://www.semantic-web-journal.net/content/facilitating-data-flows-global-publisher-using-lod2-stack

Please contact:
Christian Dirschl, Katja Eck
Wolters Kluwer Deutschland GmbH, Germany
E-mail: CDirschl@wolterskluwer.de, KEck@wolterskluwer.de

Jens Lehmann
University of Leipzig, Germany
E-mail: lehmann@informatik.uni-leipzig.de