Enabling Digital Business Transformation through an enterprise Knowledge Graph

Christian Dirschl[1], Jessica Kent[2], Jamie Schram[2], Quentin Reul[2]

[1] Wolters Kluwer Deutschland GmbH
[2] Wolters Kluwer R&D U.S. LP
{Firstname.Lastname}@wolterskluwer.com

1 Overview

Wolters Kluwer (WK) is a global provider of information, software and services for legal, tax, accounting, health, and risk and compliance professionals. WK's strategy [1] involves continuous innovation and expansion of expert solutions, including extensive use of domain knowledge, increasingly represented in the WK Knowledge Graph (KG).

KGs are a flexible knowledge representation paradigm intended to facilitate the processing of knowledge for both humans and machines. They are based on standard Semantic Web technologies and are widely regarded as a key enabler for several increasingly popular use cases, including Web search, question answering, personal assistants and enabling other AI-based applications across most industry sectors, including the legal market.

KGs are quite often generic, fragmented and incomplete, which limits their usage and coverage potential. In an industrial environment, knowledge models like KGs, controlled vocabularies and thesauri need to be combined from heterogeneous sources, mainly via mapping mechanisms. Creating an enterprise KG requires a lot of intellectual and manual knowledge work in order to end up with a scalable and sustainable result.

2 Challenges

An informed KG supports and acts as a central hub for the following four legal industry use cases (as well as many others). First, easy access to legal information across countries and languages to enhance international business efforts, e.g. in global sectors like energy, pharma or for all companies working in jurisdictions that are influenced/dominated by several jurisdictions like in the EU.

Second, better integration with standard-based legal information tools and services (e.g. EUR-LEX [2]) to accelerate LegalTech coverage, which brings added value to both companies and citizens.

Third, open (legal) data integration to enhance legal business with governmental agencies, many of which already use open data standards like EUROVOC [3].

And finally, extending legal information services to other business-oriented applications and services. For example, by adding geo information via geonames [4] to courts,

one could start socioeconomic analysis of coverage of legal advice with respect to demographics, economics and political federal structures. As per linked data principles, this could enable insights for administrations as well as for business that would otherwise not be possible.

Addressing these use cases supports WK's strategic goal to heavily expand expert solutions, like WK Germany's CaseWorx application [5], as a core means for digital business transformation [6].

The main advantages of KGs, which are their flexibility and their ability to easily aggregate large chunks of data, are also one of their main unsolved problem areas when it comes to the scalability of data. This means that it is difficult to have the information available in an easy, secure, reliable and fast way. The major challenges here are:

- Easy selection of data required in a specific project setting, e.g. better handling of multi-graph environments.
- Specific implementation of and access to universal entities that are useful in most usage contexts.
- Fast and reliable access to de-centralized KGs, both run by external as well as internal knowledge teams.
- Efficient way to query triples beyond standard SPARQL interfaces, e.g. by normalized JSON-LD usage for JSON parsers.

Flexibility also leads to the situation where the KG lacks sustainability, because a growing KG has the strong tendency to add complexity, knowledge gaps, contradictions and semantic drift over time. The major challenges here are:

- Effective data curation, including incremental updates from external sources, in order to keep the resulting KG up to date and consistent over time.
- Visualization of KG assets, so that its benefits can be made transparent to IT and business professionals. This helps to increase the usage of the KG in the end and therefore also its sustainability as part of the company's knowledge backbone.
- Creation and availability of specific domain schemas, acting as quasi-industry standards (e.g. for the construction domain within CaseWorx).
- Tools for disambiguation, e.g. company names. This requires both recognition of duplicates as well as an easy and transparent resolution.

All these challenges can and should be supported by (semi-) automated processes with high levels of insight and explanatory power, so that business-critical tasks can be leveraged by these technologies in an efficient manner.

3 Approach

The creation of a unified enterprise KG requires several components. First, unified KGs rely heavily on the adoption of a common terminology/ontology to represent the nodes and their relationships in the graph. This is, for WK, a natural use of the already existing enterprise ontology, which is both WK-specific and makes use of widely

adopted external ontologies (such as SKOS, Dublin Core, FOAF, etc.). This enables easy mapping and technology usage: for example, only minor extensions were necessary for CaseWorx (e.g. adding domain-specific properties such as "hasDefect" which models specific facts relevant to the construction domain).

Second, there need to be requirements for sources that will be used, whether they are fragmented KGs or other structured or unstructured data. The requirements that were determined to be the most important in evaluating data sources were inspired by ISO 25012 [7] and research papers on data standards (for example, "Quality Assessment for Linked Data" [8] and "Data Quality Assessment" [9]). A sample of the evaluation criteria is as follows:

- Accuracy/Reliability
  - Clear ownership – the author/website is well regarded
  - The data is commonly used
  - Data does not seem to be incorrect or inconsistent
- Relevance
  - Data or a subset of the data is specifically for the area of interest (for WK, the data is for legal, tax and accounting, or medical)
  - The data is universally useful (e.g. ISO 3166, ISO 639)
- Currency
  - Clear dates of creation/update
  - The data is updated regularly
  - The data was updated within the last 3 years (the more recent the better)
- Licensing
  - License/copyright information is easily accessible
  - License allows for commercial use
  - License is not a "share alike" license

There are several more points of evaluation WK employs which aid in our ability to quickly, accurately, and consistently determine whether to use a source. Because there is a wide variety of data in a wide variety of states, WK's evaluation criteria makes use of the MoSCoW Method [10].

Third, technology with which to build and maintain the KG must be determined. WK uses a triple store, knowledge management software, and creates programmatic transformations which support the whole Linked Data lifecycle [11]:

- Knowledge Management: Cogito Studio Express [12], PoolParty [13], VocBench [14]
- KG storage: AllegroGraph [15]
- Programmatic Transformations (proprietary and easily added to project pipelines): Conversion Services (including a program to convert Excel); Data Clean-up (removal/integration of duplicate entries, creation of persistent URIs, etc.), etc.

WK jump-started populating the KG by first adding shared controlled vocabularies/taxonomies/thesauri that are already consistently used throughout WK products. Then, outside sources (fragmented KGs, linked data, etc.), such as government

websites, shared standard vocabularies like ISO, NAICS, medical codes, and generic knowledge graphs, e.g. Wikidata, were mapped to WK's enterprise ontology and added to the KG.

Once the original mapping is completed, the process of populating the KG from a specific source can be automated and replicated, keeping the KG current and relevant.

Finally, the ability to allow for semi-open contribution in a standardized process is necessary for sustainability.

When these pieces are in place, the enterprise KG will be scalable and sustainable. WK is experienced in using Semantic Web assets and was therefore able to make use of available standards, technology, and the ongoing advocacy of contribution to/use of Semantic Web technologies already implemented in operational WK processes [16].

4 Evaluation and Lessons Learned

The approach taken is working so far. Easy inclusion of multiple sources and languages as well as acceptance and usage within the company is encouraging. However, challenges still lie in the implementation of an efficient and sustainable maintenance process. Another challenge is addressing flaws within parts of the technology used. Semantic Web Technology has made a lot of progress; however, the contextualization and usage of data are still major challenges. For example, the contextualization of CaseWorx requires disambiguation (e.g. company names) and semantic integration (e.g. each mandate is represented as a subgraph within CaseWorx; building different views on top of KGs). Both aspects are aided by the enterprise KG, but there is still work that needs to be done. Similarly, data usage is a major concern for CaseWorx as public ontologies are still very generic and may have licensing or governance issues and need to comply with customer privacy needs (e.g. each customer only sees his own data). Licensing is very important, since many open sources are only available under a ShareAlike copyright [17], which prevents real business usage. Dual licensing models would be highly appreciated.

We recommend adhering to the following basic rules: a) use either stable vocabularies or mature vocabularies with associated governance and documented maintenance processes; b) establish clear rules for KG integration beyond being mapped to the ontology; c) use Named Graphs to make maintenance easier.

5 Conclusion and Future Work

KGs have the potential to support and enable WK's strategic goals. Already existing fragments are available for general usage (e.g. within CaseWorx), and first extensions have been made. Use case and business impact analysis is on its way. We are currently adding new sub-graphs and are working towards sustainable mapping mechanisms with stable and scalable maintenance and development processes in the existing WK ecosystem. Further collaboration with the scientific community is needed.

References

1. Wolters Kluwer Homepage, https://wolterskluwer.com/company/about-us/strategy.html, last accessed 2020/03/04.
2. EUR-LEX Homepage, https://eur-lex.europa.eu/, last accessed 2020/04/30.
3. EUROVOC Homepage, https://eur-lex.europa.eu/browse/eurovoc.html, last accessed 2020/04/30.
4. CaseWorx Homepage, https://www.caseworx-baurecht.de/, last accessed 2020/04/30.
5. Geonames Homepage, https://www.geonames.org/, last accessed 2020/04/30.
6. Pellegrini, T., Dirschl, C., Eck, K.: Linked data business cube: a systematic approach to semantic web business models; In: AcademicMindTrek '14 Proceedings of the 18th International Academic MindTrek Conference: Media Business, Management, Content & Services, pp 132-141; ACM (2014).
7. ISO/IEC 25012:2008 page, https://www.iso.org/standard/35736.html, last accessed 2020/04/30
8. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2015). Quality assessment for Linked Data: A Survey. Semantic Web, 7(1), 63–93. doi: 10.3233/sw-150175
9. Pipino, L.L., Lee, Y.W., & Wang, R.Y. (2002). Data quality assessment. Communications of The ACM, 45, 211-218.
10. Wikipedia page: “Moscow Method”, https://en.wikipedia.org/wiki/MoSCoW_method, last accessed 2020/04/30
11. Auer, S., Bühmann, L., Dirschl, C., Erling, O., Hausenblas, M., Isele, R., & Williams, H. Managing the life-cycle of Linked Data with the LOD2 Stack. In: The Semantic Web–ISWC 2012, pp 1-16; Springer, Berlin Heidelberg (2012).
12. Cogito Studio Homepage, https://expertsystem.com/products/cogito-studio/, last accessed 2020/04/30.
13. PoolParty Semantic Suite Homepage, https://www.poolparty.biz/, last accessed 2020/04/30.
14. VocBench Homepage, http://vocbench.uniroma2.it/, last accessed 2020/04/30.
15. Allegrograph Homepage, https://franz.com/agraph/allegrograph/, last accessed 2020/04/30.
16. Hondros, C., http://linkeddatadeveloper.com/Projects/Linking-Enterprise-Data/Manuscript/led-hondros.html, last accessed 2020/04/30.
17. Creative Commons Homepage, https://creativecommons.org/licenses/by-sa/2.5/, last accessed 2020/04/30.