www.jusletter-it.eu

Christian Dirschl

Thesaurus Generation and Usage at Wolters Kluwer Deutschland GmbH

This paper describes the creation and maintenance process of legal thesauri at Wolters Kluwer Germany. In addition, thesaurus usage within the innovative legal platform JURION and beyond is covered. Assertions on general strategic decisions on content and data acquisition, processing and maintenance put this effort in a more general context. A section on best practices and concrete recommendations mainly for thesaurus creation and maintenance, followed by conclusions and future work complete this overview.

Category: Articles
Region: Germany
Field of law: Law and Language
Collection: Conference Proceedings IRIS 2016

Citation: Christian Dirschl, Thesaurus Generation and Usage at Wolters Kluwer Deutschland GmbH, in: Jusletter IT IRIS

Contents

1. Wolters Kluwer Deutschland GmbH and JURION

[Rz 1] This section introduces Wolters Kluwer and its legal knowledge platform JURION and gives some context information on the current business as a whole.

1.1. Wolters Kluwer

[Rz 2] Wolters Kluwer Deutschland GmbH is an information services company specializing in the legal, business and tax sectors. Wolters Kluwer provides pertinent information to professionals in the form of literature, software and services. Headquartered in Cologne, it has over 1,200 employees located at over 20 offices throughout Germany, and has been conducting business on the German market for over 25 years.

[Rz 3] Wolters Kluwer Germany is part of the leading international information services company, Wolters Kluwer n.v., located in Alphen aan den Rijn (The Netherlands). The core market segments, targeting an audience of professional users, are legal, business, tax, accounting, corporate and finance services, and healthcare. Its shares are quoted on the Euronext Amsterdam (WKL), and are included in the AEX and the Euronext 100 indices. Wolters Kluwer has annual sales of €3.6 billion (2014), employs approximately 19,000 people worldwide and has over 40 offices located throughout Europe, North America, Asia Pacific region and in Latin America.

1.2. JURION

[Rz 4] JURION is the legal knowledge platform developed by Wolters Kluwer Germany. It is not only a legal search platform, but considers search for legal information as an integrated part of the lawyer's daily processes. JURION combines competencies in the areas of legal publishing, software, portal technology and services, which cover all core processes of the lawyer within one single environment by connecting and integrating many different internal and external data sources.

2. Thesaurus Creation, Maintenance and Usage

[Rz 5] This is the main section of this paper, describing how legal thesauri are created, enriched, maintained and used. Also some information on effort and costs is included.

[Rz 6] A thesaurus as described in this paper is following the definition ISO 25964[1], so our schema is represented in SKOS[2] . This is important for several reasons. First, the standard gives already guidance towards structure and coverage. Second, sticking to standards avoids dependencies on specific tools and interfaces, so that technological decisions in the beginning can later on easily be adapted or even withdrawn.

2.1. Thesaurus Creation

[Rz 7] Thesaurus creation is a very challenging task, especially when it has to be created from scratch. We have developed a standardized process, which ensures a proper balance between exhaustiveness and maintenance.

[Rz 8] The good news here is that our experience and also our existing folio and digital products give us a good head start, so we do not really start from scratch.

[Rz 9] Table of contents in legal handbooks already use a well-defined topical structure that reflects also the thinking of the end-users of the thesaurus information, e.g. on a legal search platform. Keyword indexes also available in books add another layer of granularity to the structure and offer first candidates for synonyms and cross-references. Legal citations in texts can also be exploited for e.g. cross-references and links to additional important sources.

[Rz 10] So the standard process we have implemented looks in general like:

- Choose around five standard handbooks that properly cover your domain
- Extract the table of contents and create a taxonomical backbone of around 800 concepts, which are structured not more than three levels deep
- Enrich this basic thesaurus with keywords from the index
- Enrich this thesaurus further with common keywords used by users in your search platform
- Add cross references between concepts by taking into account law structures reflected in the legal references
- End up with a first version of a thesaurus with maximum 1.500 concepts and three levels deep (with exceptions to four levels where necessary)

[Rz 11] Since we have a standard process in place and since we can build on a lot of valuable material that we can re-use within the creation process, we calculate respective costs and efforts for one legal domain like IP law or construction law with:

- 10 to 20k€external costs for thesaurus creation
- 1 to 2 person months internal effort for data preparation, quality assurance and thesaurus implementation in our internal processes

1 http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=53657.

2 http://www.w3.org/2004/02/skos/.

2.2. Thesaurus Enrichment

[Rz 12] Legal thesauri are the backbone of many application features in JURION, e.g. faceted search or concept search. In order to leverage these capabilities, we have added distinct properties and relationships to SKOS standard (e.g. certain legal relationship types like «one court decision overrules another court decision»). In addition, we have mapped and linked our thesauri to other standard thesauri like EUROVOC[3] or STW[4] in order to further exploit existing knowledge that is publicly available and in order to support the initiative to achieve a common data ecosystem across Europe.

2.3. Thesaurus Maintenance

[Rz 13] Thesaurus maintenance is a challenge. Normally there are no processes for that in place within an organization. So there are two main options: either to maintain the thesaurus externally or to establish such an internal process. Both options are possible, but the more relevant issue in this respect is how requests for changes are collected and how the decision process for implementing these changes is governed. Since we have a very sophisticated process in place for maintaining our XML content structures, we decided to apply this process to knowledge models as well. This also means that there is a versioning and release process normally based on a six month release cycle. As with the creation process, also the maintenance process is very conservative, accepting proposed changes only when a clear business need is identified and documented (!).

2.4. Thesaurus Usage

[Rz 14] As already mentioned, thesaurus information is reflecting legal domain knowledge in a machine-readable and processable way and therefore enables a wide variety of relevant use case scenarios [Dirschl/Eck 2015, Dirschl 2014, Blumauer/Dirschl 2013]. This starts with functionality, where the thesaurus is visible for the customer like using the hierarchical backbone for faceting browsing [Lee/Kim/Seo/Kim/Lee/Jung/Dirschl 2011] or using a visual mapper tool[5] for exploratory search. In addition, thesaurus information can be used for a better relevance ranking of search results as well as for disambiguation of keyword searches. It can also support the very powerful autosuggest feature, where searches are proposed by the system whilst the user is keying in his search terms.

3. Standardized content and data processing

[Rz 15] The pure existence and usage of thesaurus information in digital products does not exploit its full potential. It needs to be part of a business strategy that is based on the notion that information always needs to be contextualized and smart. Therefore, we have created a standardized content and data processing pipeline, using standards and open source technology based

3 http://eurovoc.europa.eu/drupal/?q=node.

4 http://zbw.eu/stw/version/latest/about.en.html.

5 https://www.youtube.com/watch?v=28HDqsDVnvk.

on semantic web principles wherever possible [Dirschl/Pellegrini/Nagy/Eck/Van Nuffelen/Ermilov 2014, Auer/Bühmann/Dirschl/Erling/Hausenblas/Isele/Williams 2012]. Figure 1 shows the content workflow as well as the semantic search pipeline.

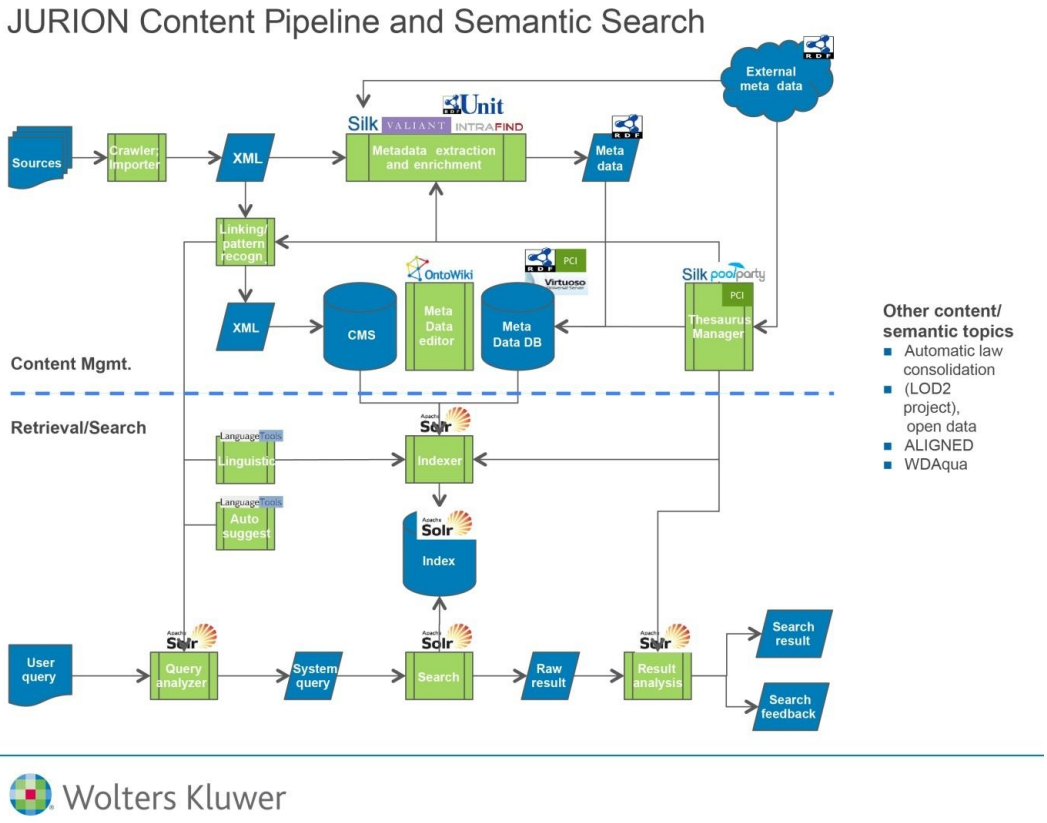

Figure 1: JURION Content Pipeline

[Rz 16] Within the content pipeline, metadata is extracted from the proprietary WKD XML schema and transformed in RDF. Due to regular changes in the XML format, the correct transformation process based on existing XSLT scripts must be secured, so that no inconsistent data is fueled into the metadata database [Dirschl/Eck/Lehmann 2014, Dirschl, Eck 2014].

[Rz 17] In the Thesaurus Manager controlled vocabularies and domain models are created, maintained and delivered for further usage, based on SKOS standard. The integrity of the knowledge management system as a whole also needs to be ensured. Therefore, regular local and global quality checks need to be executed, so that e.g. inconsistencies across different controlled vocabularies can be detected and resumed.

[Rz 18] Data-intensive systems like JURION are highly dependent on the metadata that steers many of their core functionalities. The indexing process of a search engine includes more and more additional information on top of the pure text. Queries are analyzed for legal references and keywords which are matched against existing data in the metadata management systems. Once there are matches, the semantic relations are shown in the results overview by specific references to texts and related knowledge visualizations.

[Rz 19] Extensive and smart quality checks on content based on the requirements derived from this process could strengthen the stability of the overall system.

4. Best practices

[Rz 20] This paper is focusing on how thesauri are gaining importance in knowledge-intensive companies. We have experience on that for more than ten years and therefore we can give some suggestions on what to be aware of and how to proceed in thesaurus usage.

[Rz 21] Based on our experience, there is normally no need to really start from scratch. A lot of structural knowledge and existing knowledge models that at least partly cover the task at hand are hidden on the web or available in different publications. It is always worth having a look and get inspiration. Also get in touch with dedicated experts in that field. And although the legal domain is a very specific one, the main tasks that need to be tackled are the same in other domains. So look at industries like health, where there exists a lot of experience for more than twenty years. But: do not rely on general language resources that are not adapted to «legalese». They will not only have huge gaps concerning legal topics, they will also too often be misleading, since even everyday terms have specific semantics in a legal environment, which these vocabularies usually do not reflect.

[Rz 22] Go for automatic support wherever possible. But be aware of the fact that currently no tool available on the market can do this job in an unsupervised way with acceptable quality. So go for a semi-automatic approach with a lot of intellectual quality assurance cycles right from the start. The main challenge here is to find a subject matter expert, who has the required legal knowledge and is capable of applying this to more technical structures like thesauri.

[Rz 23] Find a proper pragmatic balance between thesaurus coverage and maintenance costs. We decided that we do not aim for having one single thesaurus in place that is covering everything, but to have smaller, domain specific thesauri, that are much easier to handle. Mapping technologies are used to bridge the gap between thesauri wherever this is desirable, e.g. documenting the fact that there are notions around «contracts» in many different areas of law. As a rule of thumb, we have decided to go for a thesaurus with 1500 to 2000 concepts with a maximum depth of four levels.

[Rz 24] Use standards as much as possible, both from the modelling point of view as well as from the tooling point of view. This will ensure that your effort is preserved over time and it enables you to make your data interoperable with other data, which is a capability that will gain more and more importance in the coming years. Also consider to at least make parts of your knowledge models publicly available using open licenses, so that you can benefit from the emerging web of data [Pellegrini/Dirschl/Eck 2014].

5. Conclusion and future work

[Rz 25] In this paper we have shown how thesaurus creation and maintenance is part of our daily business at Wolters Kluwer Deutschland GmbH. We have described that thesauri are an important knowledge source that needs to be embedded in a more global business and content strategy, which also includes a proper content pipeline based on standards.

[Rz 26] We are still working on extending the number of our thesauri and we continuously refine our existing ones. We are currently looking into how language technology can help us to transfer our monolingual thesauri into multilingual thesauri, which would extend their applicability e.g.

towards cross-border applications. With the help of the ALIGNED project[6] we are also working on further streamlining our content pipeline, but also on finding out how a better integration of content development and software engineering could look like in the coming years.

6. Acknowledgement

[Rz 27] This work has received funding from the European Union's Horizon 2020 re-search and innovation program under grant agreement No 644055, the ALIGNED project (www.aligned-project.eu).

7. Literature

Auer, Sören/Bühmann, Lorenz/Dirschl, Christian/Erling, Ori/Hausenblas, Michael/Isele, Robert/Williams, Hugh, Managing the life-cycle of Linked Data with the LOD2 Stack. In: The Semantic Web–ISWC 2012, Springer Berlin Heidelberg 2012, pp 1–16.

Blumauer, Andreas/Dirschl, Christian, Linked Data – Das Ende des Dokuments?. In: DOK Magazin 6/13, Verlag Marketing Projekt 2000 GmbH 2013, pp 12–16.

Dirschl, Christian, Linked Data – The End of the Document?. In: Wissen verändert – Beiträge zu den Kremser Wissensmanagement-Tagen 2014, Edition Donau-Universität Krems 2015, pp 41–48.

Dirschl, Christian/Eck, Katja, Verlage müssen sich neu erfinden. In: Corporate Semantic Web, Springer Berlin Heidelberg 2015, pp 129–144.

Dirschl, Christian/Eck, Katja, Linked Data als integraler Bestandteil der Kernprozesse bei Wolters Kluwer Deutschland GmbH. In: Linked Enterprise Data, Springer Berlin Heidelberg 2014, pp 289–295.

Dirschl, Christian/Eck, Katja/Lehmann, Jens, Supporting the Data Lifecycle at a Global Publisher using the Linked Data Stack. In: ERCIM News 96, ERCIM EEIG 2014, pp 38–39.

Dirschl, Christian/Pellegrini, Tassilo/Nagy, Helmut/Eck, Katja/Van Nuffelen, Bert/Ermilov, Ivan, LOD2 for Media and Publishing. In: Linked Open Data – Creating Knowledge out of Interlinked Data, Springer Lecture Notes in Computer Science Volume 8661, Springer Berlin Heidelberg 2014, pp 133–154.

Lee, S./Kim, P./Seo, D./Kim, J./Lee, J./Jung, H./Dirschl, Christian, Multi-faceted Navigation of Legal Documents. In: 2011 International Conference on and 4th International Conference on Cyber, Physical and Social Computing, IEEE Computer Society Washington, DC 2011, pp 537–540.

Pellegrini, Tassilo/Dirschl, Christian/Eck, Katja, 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, ACM 2014, pp 132–141 (Winner of Best Paper Award).

6 http://aligned-project.eu/.