LawSampo Portal and Data Service for Publishing and Using Legislation and Case Law as Linked Open Data on the Semantic Web

Eero Hyvönen[1,2], Minna Tamper[1,2], Esko Ikkala[1], Mikko Koho[1], Rafael Leal[1], Joonas Kesäniemi[1], Arttu Oksanen[1], Jouni Tuominen[1,2] and Aki Hietanen[3]

[1]*Semantic Computing Research Group (SeCo), Aalto University, Finland*
[2]*Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, Finland*
[3]*Ministry of Justice, Finland, Finland*

Abstract
This paper argues for the idea of publishing legislation and case law as Linked Open Data (LOD) on the Semantic Web, to cater several user groups, including the general public, legislators, lawyers, researchers of legal informatics, and application developers. To support the argument, the proof-of-concept system *LawSampo – Finnish Legislation and Case Law on the Semantic Web* is introduced, including a semantic portal and a LOD service. Based on the Sampo Model, the main novelty of LawSampo is the provision of heterogenous distributed legal data through multiple application perspectives for faceted searching and exploring the data and for data analysis in legal informatics.

Keywords
Linked data, Case law, Legislation, Semantic portal

1. Introduction

Legislation and case law are widely published online by governments to make jurisdiction transparent and freely accessible to the public, organizations, and lawyers [1]. The Web provides a promising medium for publishing such big data. There are, e.g., portals, such as legislation.gov.uk for the legislation for the UK, Scotland, Wales, and Northern Ireland[1], and EU level systems, such as the EU Cellar[2] and the ECLI Search Engine[3] for the case law.

However, legal documents are often available only as texts for the humans to read with little metadata available, which makes them hard to use in applications of legal informatics[4] [2], e.g., in computational law[5]. To address the problem, this paper argues that legislation and case law

International Workshop on Artificial Intelligence Technologies for Legal Documents (AI4LEGAL), virtual, 23 October 2022, Hangzhou, China
0000-0003-1695-5840 (E. Hyvönen); 0000-0002-3301-1705 (M. Tamper); 0000-0002-9571-7260 (E. Ikkala); 0000-0002-7373-9338 (M. Koho); 0000-0001-7266-2036 (R. Leal); 0000-0002-3770-0006 (J. Kesäniemi); 0000-0003-2327-6942 (A. Oksanen); 0000-0003-4789-5676 (J. Tuominen)



CEUR Workshop Proceedings (CEUR-WS.org)

[1]https://www.legislation.gov.uk
[2]https://data.europa.eu/euodp/en/data/dataset/sparql-cellar-of-the-publications-office
[3]https://e-justice.europa.eu/content_ecli_search_engine-430-en.do
[4]https://en.wikipedia.org/wiki/Legal_informatics
[5]https://law.stanford.edu/2021/03/10/what-is-computational-law/

should be published and used as Linked Open Data (LOD) on the Semantic Web. To support the argument, a case study based on Finnish legislation and case law is overviewed as the system LawSampo – *Finnish Legislation and Case Law on the Semantic Web* that consists of a LOD service and a semantic portal, extending our earlier short paper [3]. LawSampo is based on the more general *Sampo Model* [4] for collaborative LOD publication that has been applied to a series of portals[6] in Digital Humanities.

In the following, we first describe the LOD underlying LawSampo. After this using the portal and data service are explained. In conclusion, related works are discussed and contributions and lessons learned are summarized.

2. LawSampo Linked Open Data

Primary Data Finnish legislation and case law decisions have been published as web documents since 1997 in the Finlex Data Bank[7]. Although this service is widely used, it does not provide machine-readable legal information as open data. To address this, we published a selection of Finlex data as the Semantic Finlex [5] LOD service that currently contains ca. 28 million triples. We transformed this data into a simplified data model suitable for the portal, and the data was enriched by data linking and knowledge extraction techniques.

Data Model LawSampo represents legislation and case law using a simple data model. The legislation data consists of statutes and their sections, whereas the case law data includes court decisions with language versions. Metadata about the instances are given using various classes and properties, mostly aligned with DCMI Metadata Terms[8]. The data model schema is available and documented at the namespace URI http://ldf.fi/schema/lawsampo/.

Data Transformation The LawSampo data transformation process is presented in Fig. 1. Semantic Finlex data is first transformed and filtered with SPARQL Construct queries. The Legislation RDF data contains only the latest versions of the consolidated legislation. Next, keyword extraction and document classification is employed to the textual contents to link them to corresponding subject keywords and life situations, respectively. After this, the entities are further linked to facet ontologies of time and EU legislation—The LawSampo portal is based on faceted search. The third step involves applying Named Entity Linking to the textual contents of Legislation and Case Law RDF. The facet ontologies are transformed from CSV format into RDF.

Internal linking The data was linked internally to improve the references to other documents. The links to legal documents needed more processing as the statutes for instance may refer to more concrete part of the statute in a specific version. Unfortunately, the Semantic Finlex data is not complete and requires some human interpretation. A challenge in the data is that the court decisions only have the judgment date but not the dates for the events that are under investigation in the document. However, the judgment is based on legislation that was valid during the time of the events that are being evaluated. Therefore, the linking from decisions to statutes cannot be done to a specific statute version but only to the current consolidated version

[6] https://seco.cs.aalto.fi/applications/sampo/
[7] http://www.finlex.fi
[8] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

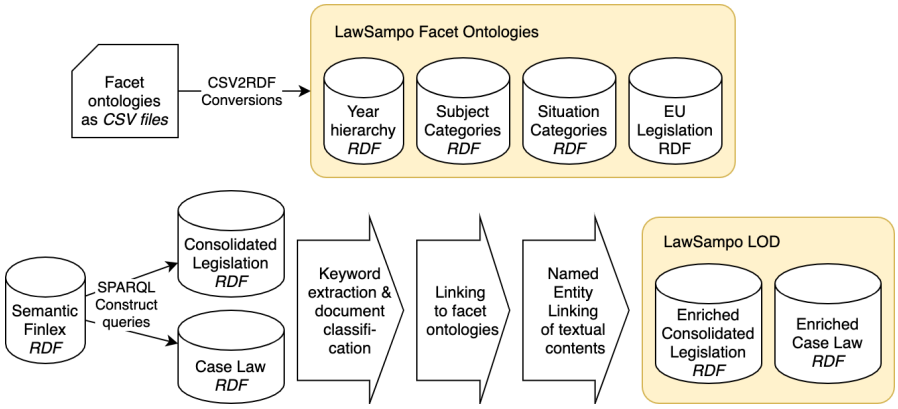

Figure 1: LawSampo data transformation process from Semantic Finlex

of it that doesn't take into account what version of the statute was in force at the time of the judged event.

External linking The LawSampo dataset has been linked to external data sources including EU Cellar and the Finlex service. The EU Cellar links have been extracted from the original Finlex statute documents [5] and included in the LawSampo data with their descriptive texts. The statutes and the court decisions were linked to the Finlex service to show the original source data.

Terminological linking The legal terms occurring in the texts were linked to their explanations in order to make the texts more readable to the layman by a contextual reader [6]. The Nelli [7] and ARPA tools [8] are used to identify term instances in legal documents and link them to vocabularies with terminological definitions: the Combined Legal Concept Ontology [9], Finnish DBpedia, and the Helsinki Term Bank for the Arts and Sciences[9]. A term instance of a document contains information about its literal representation, links to its definitions in the vocabularies and the document, and a count telling how many times the term in mentioned in the document. The count information is needed for generating tag clouds summarizing the contents of the documents. The instances are represented in RDF and added to the LawSampo dataset.

Keyword extraction and document classification The subject indexing tool Annif[10] [10], developed by the National Library of Finland, is used to perform keyword extraction for all the legislative documents. Annif is capable of using different algorithms in order to return suggested keywords and their respective weights for a given input text. The developers also provide a REST API containing various pre-trained projects that combine different algorithms. LawSampo uses the *yso-fi* pre-trained project, which integrates *TF-IDF*, *Maui* and *Parabel*. The first two are *lexical* algorithms, so they directly match terms to a vocabulary, whereas Parabel uses an *associative approach* that is able to also find indirect correlations between words [10]. This mix provides results that are not only grounded on the text of the documents but also able

[9] https://tieteentermipankki.fi
[10] https://annif.org

to extrapolate their specific wording. *Yso-fi* is trained on bibliographical metadata from Finnish museums, archives and libraries. Since the training data is labeled with terms from the General Finnish Ontology (YSO)[11], the API returns keywords identified by unique YSO URIs.

A zero-shot classification system based on the extracted keywords [11] is also used. It works by first transforming the documents into vectorial representations via the word-embedding algorithm *fastText* [12], using a pre-trained Finnish language model offered by the fastText developers [13]. The document representations are then calculated as the average embedding of their respective keywords. A similar treatment involving Annif and fastText is given to a list of category labels representing different life situations, such as *asuminen, kiinteistö* 'housing, real state', *ihmisoikeudet, perusoikeudet* 'human rights, basic rights' or *omaisuus, kaupankäynti, kuluttajansuoja* 'property, commerce, consumer protection'. This results in vectorial representations for each category as well. The classification is then carried out by comparing the document vectors with the category vectors via cosine similarity. Each document is assigned the 5 best-fitting categories whose weight is within 95% of their top category.

Keywords and categories are used in two different ways in LawSampo: they are used to label the respective legislative documents as metadata (respectively as *subject keywords* and *life situation/topic*), and they also form the basis for a semantic search system which will be explained later in this paper.

Linked Open Data Service The LawSampo data service adopts the 5-star Linked Data model[12], extended with two more stars, as suggested in the Linked Data Finland model and platform [14]. The 6th star is obtained by providing the dataset schemas and documenting them. The LawSampo schema can be downloaded from the service[13] and the data model is documented using the LODE service[14]. The 7th star is achieved by validating the data against the documented schemas to prevent errors in the published data. LawSampo attempts to obtain the 7th star by applying different means of combing out errors in the data within the data conversion process. The LawSampo data model and its integrity constraints are presented in a machine-processable format using the ShEx Shape Expressions language[15] [15]. We have made initial validation experiments with the PyShEx[16] validator. Based on the experiments, we have identified errors both in the schema and the data, and a full-scale ShEx validation phase for the data conversion is underway.

The Linked Data service is powered by the Linked Data Finland[17] publishing platform that along with a variety of different datasets provides tools and services to facilitate publishing and re-using Linked Data. All URIs are dereferenceable and support content negotiation by using HTTP 303 redirects. The data is available as an open SPARQL endpoint[18]. As the triplestore, Apache Jena Fuseki[19] is used as a Docker container, which allows efficient provisioning of

[11] https://finto.fi/yso/en/
[12] https://www.w3.org/DesignIssues/LinkedData.html
[13] https://www.ldf.fi/dataset/lawsampo
[14] https://essepuntato.it/lode/
[15] https://shex.io
[16] https://github.com/hsolbrig/PyShEx
[17] http://ldf.fi
[18] https://ldf.fi/lawsampo/sparql
[19] https://jena.apache.org/documentation/fuseki2/

resources (CPU, memory), portability, and scaling. Varnish Cache web application accelerator[20] is used for routing URIs, content negotiation, and caching.

3. Using LawSampo Portal

This section overviews how the LawSampo portal and the underlying LOD service are used in practise. The system is based on the Sampo Model [4] that is an informal collection of six principles for 1) LOD publishing and 2) designing semantic portal user interfaces (UI), supported by the Sampo-UI framework [16].

The landing page of the LawSampo portal depicted in Fig. 2 offers five different application perspectives. Semantic faceted semantic search is used for filtering data of interest out after which the data can be either browsed or analyzed using a set of seamlessly integrated data-analytic tools. The five application perspectives are explained below.

Figure 2: LawSampo landing page with five application perspectives

1. Statutes Perspective By clicking on the Statutes perspective box, a faceted search interface for searching and browsing statutes is opened. The facets on the left include document type (with seven subtypes), statute type, year, and related EU regulation. After filtering out a set of documents (or a particular document) of interest, the user is provided with two options. First, the user can select a document from the result list and a "homepage" of the document opens, showing not only the document but also linked contextual information related to it such as the referred EU regulations linked to EU Cellar or other documents from Semantic Finlex referring to it. The LawSampo application utilizes enriched data and shows annotated statute documents. The annotations are highlighted in the text and by hovering over the annotation, the user can see the explanation to the term and links to external portals such as Wikipedia or the Helsinki Term Bank for the Arts and Sciences to learn more about the term. The terms are also used to create tag cloud visualizations to give the user an idea what the text is about.

[20] https://varnish-cache.org

2. Sections Perspective The Sections perspective operates in the same way, but here it is possible to search and explore consolidated legislation on a more focused section level.

3. Case Law Perspective In the Case Law perspective, a similar faceted search interface opens for searching and browsing court decisions. In this case, the facets include court, judge, and keywords characterizing the subject matter of the judgment. Similarly to statutes, the case law view shows the results based on the facet selections as a list for the user. From this point on the user can view the court decision details at its "homepage". Similarly to statutes, the court decision's page contains the annotated text document, a tag cloud, and more related information about it. The court decisions also have been enriched in the portal with related case law documents based on Semantic Finlex Case Law Finder. It retrieves documents that are textually similar to the selected document and the results are listed in the table tab.

In addition to the court decision listing and homepages, the user's choices also influence the other tabs in the case law perspective, for example, the statistics, such as the facet's pie charts or the by year bar charts for the court decisions. By selecting a value from a facet, all other facets and results update and the distribution of court decisions by year or by facet (e.g., by court, in the court facet) show the updated results. With these statistical tools the user can view and study the case law data. Fig. 3 is a screenshot from the Case Law perspective's plot depicting the number of court decisions by year. The plot shows the number of court decisions with the judgment date information on a timeline.

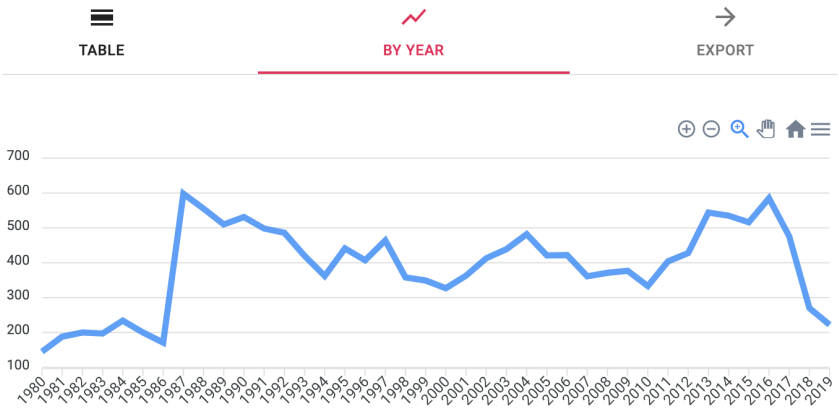

Figure 3: Number of court decisions in 1990–2019 in Case Law Perspective

The Case Law perspective also enables the user to export the faceted search SPARQL query into the Yasgui[21] tool on a separate tab EXPORT. The data can then be explored further by editing the SPARQL query, and the results can be downloaded for further study, e.g., into a spreadsheet program in CSV form.

4. Contextual Search Perspective The fourth perspective, named Contextual Search, allows for searching legal documents based on the end user's life situation at hand (e.g., divorce). This search system employs the Relevance Feedback Search (RFBS) paradigm, which works

[21] https://yasgui.triply.cc/

by refining the search parameters iteratively, with input from the user, in order to improve its results. In LawSampo this is done by offering the user keyword and category suggestions based on the results of the previous search round: by activating or deactivating these suggestions, the user gradually redirects the query towards a more satisfying result. This kind of search is argued to be useful in situations where it is difficult for the user to formulate a traditional search query. This application is described in [11] in more detail.

5. Similarity-based Case Law Search The landing page also provides a link to an application for searching court decisions, based on the assumption that textually similar cases are relevant for the information need. The user is able to input a text document as a query, either by uploading a file or by writing text directly into the form. This application can be helpful when, e.g., someone has received a court decision and is interested to see whether the verdict is fair in comparison with other similar cases. Hopefully, such possibility could lessen unnecessary appeals to higher courts in the future. In this application several methods for finding similar cases were tested when implementing this application including TF-IDF, Latent Dirichlet Allocation, Word2Vec, and Doc2Vec [17].

In addition to the ready-to-use statistical applications integrated into the LawSampo portal, the underlying open SPARQL endpoint can be used for querying, analyzing, and visualizing the data in flexible ways and external tools. For example, the visualization in Fig. 4 shows how the number of court decisions of different kinds in the data change in time during 1980–2020. This graph has been created using Jupyter notebooks. The figure shows how the number of civilian and criminal cases decreases in time, and that the cases of the Administrative court dominate the dataset. There is also a number of court decisions without type of the matter described. Also their number is decreasing but slowly.

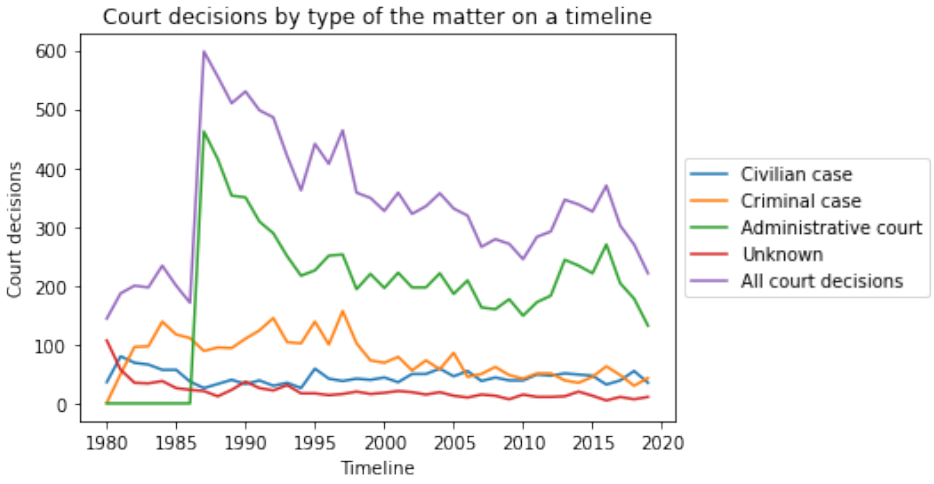

Figure 4: Number of court decisions by type of the matter

4. Discussion

Related Works Our work on legal Linked Data services was influenced by the MetaLex Document Server[22] [18] and related national online services for legal documents in Greece, Luxemburg[23], France, Norway[24], and the U.S. [19]. EU Cellar publishes EU legislation as LOD. Companies provide legal services for searching and exploring legislation and case law, and Google Scholar has a specific search application for cases in the various courts of the states[25] in the U.S.

Contributions and Challenges This paper applied the Sampo Model, developed originally for Digital Humanities research, to a novel use case in legal informatics. Legislation and case law data are provided through multiple end user groups and purposes through application perspectives. The documents are automatically enriched with contextual linked data, and the end user is provided with ready-to-use faceted search and data-analytic tooling for analyzing the documents and their relations.

However, extracting and linking references of legal documents requires still more work. The references to legal documents can be made in various ways and the labels we currently have in our databases are not enough to identify all the ways in which the references are made in texts. There are references made using the official names or nick names that exist in the Finlex database, but some references are made with unidentified acronyms or by twisting the order of words in the names, which may produce unidentifiable wordings for different statute names. It would be much easier to add metadata about related documents manually when indexing the documents than trying to extract the links from unstructured texts afterwards. The biggest semantic challenge we encountered in our work was that the statutes are not stable but their sections are dynamically added, cancelled, and modified in time by other statutes. In the Finnish legislation system, systematic time series of consolidated versions of legislation are not available, but only the initial versions of the statuses and series of changes made to them afterwards. The court decisions are, however, always made based on the legislation in force at the time of the judicial offense, which makes the linking between legislation and case law difficult. LawSampo has access only to the latest versions of manually consolidated statues available in Finlex, and the problem of finding out how the statutes may have changed in time is left to the end user. In many cases, the court decision does not even tell when the judged offence was made, but may only refer to a lower court decision where the date information may be available. From data publishing point of view this information should be added to the decision metadata already at the courts.

Usability of the LawSampo Portal has not been evaluated yet. However, the Sampo model has been evaluated in some other Sampo portals [20] suggesting feasibility of the model in general. An empirical evidence of this is also that Sampo portals are widely used on the Web by up to millions of users [4].

In spite of the challenges and complexities of the underlying data, we are confident that that proposed LOD approach is feasible and usable in practice, and plan to make the LawSampo

[22] http://doc.metalex.eu
[23] http://legilux.public.lu/editorial/eli
[24] http://lovdata.no/eli
[25] https://scholar.google.com/scholar_courts

prototype publicly available in the future.

Acknowledgments We thank Tiina Husso, Risto Talo and Jari Linhala for collaborations. Funding was provided by the Ministry of Finance, the Academy of Finland, the EU project InTaVia[26], and action Nexus Linguarum[27] on linguistic data science. CSC – IT Center for Science provided computational resources.

References

[1] M. van Opijnen, G. Peruginelli, E. Kefali, M. Palmirani, Online publication of court decisions in europe, Legal Information Management 17 (2017) 136—-145. doi:10.1017/S1472669617000299.

[2] S. Erdelez, S. O'Hare, Legal informatics: Application of information technology in law, Annual Review of Information Science and Technology 32 (1997).

[3] E. Hyvönen, M. Tamper, A. Oksanen, E. Ikkala, S. Sarsa, J. Tuominen, A. Hietanen, LawSampo: A semantic portal on a linked open data service for Finnish legislation and case law, in: The Semantic Web: ESWC 2020 Satellite Events. Revised Selected Papers, Springer, 2019, pp. 110–114.

[4] E. Hyvönen, Digital humanities on the Semantic Web: Sampo model and portal series, Semantic Web – Interoperability, Usability, Applicability (2022) 1–16. URL: https://doi.org/10.3233/SW-223034.

[5] A. Oksanen, J. Tuominen, E. Mäkelä, M. Tamper, A. Hietanen, E. Hyvönen, Semantic Finlex: Transforming, publishing, and using Finnish legislation and case law as linked open data on the web, in: Knowledge of the Law in the Big Data Age, IOS Press, 2019, pp. 212–228.

[6] E. Mäkelä, T. Lindquist, E. Hyvönen, CORE - a contextual reader based on linked data, in: Proceedings of Digital Humanities 2016, long papers, 2016, pp. 267–269. URL: http://dh2016.adho.org/abstracts/2580.

[7] M. Tamper, A. Oksanen, J. Tuominen, A. Hietanen, E. Hyvönen, Automatic annotation service APPI: Named entity linking in legal domain, in: The Semantic Web: ESWC 2020 Satellite Events, volume 12124 of *Lecture Notes in Computer Science*, Springer-Verlag, 2020, pp. 208–213. doi:10.1007/978-3-030-62327-2_36.

[8] E. Mäkelä, Combining a REST lexical analysis web service with SPARQL for mashup semantic annotation from text, in: Proceedings of the ESWC 2014 demonstration track, Springer, 2014, pp. 424–428.

[9] M. Frosterus, J. Tuominen, E. Hyvönen, Facilitating re-use of legal data in applications – Finnish law as a linked open data service, in: Proceedings of the 27th International Conference on Legal Knowledge and Information Systems (JURIX 2014), IOS Press, 2014, pp. 115–124.

[10] O. Suominen, Annif: DIY automated subject indexing using multiple algorithms, LIBER Quarterly 29 (2019) 1–25. doi:10.18352/lq.10285.

[11] R. Leal, J. Kesäniemi, M. Koho, E. Hyvönen, Relevance Feedback Search Based on Auto-

[26] https://intavia.eu/

[27] https://nexuslinguarum.eu/the-action

matic Annotation and Classification of Texts, in: 3rd Conference on Language, Data and Knowledge (LDK 2021), volume 93 of *Open Access Series in Informatics (OASIcs)*, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2021, pp. 18:1–18:15. doi:10.4230/OASIcs.LDK.2021.18.

[12] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching Word Vectors with Subword Information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146. doi:10.1162/tacl_a_00051.

[13] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages, in: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[14] E. Hyvönen, J. Tuominen, M. Alonen, E. Mäkelä, Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets, in: ESWC 2014 Satellite Events, Springer, 2014, pp. 226–230. doi:10.1007/978-3-319-11955-7_24.

[15] K. Thornton, H. Solbrig, G. S. Stupp, J. E. L. Gayo, D. Mietchen, E. Prud'hommeaux, A. Waagmeester, Using shape expressions (ShEx) to share RDF data models and to guide curation with rigorous validation, in: The Semantic Web. ESWC 2019, Springer, 2019, pp. 606–620. doi:10.1007/978-3-030-21348-0_39.

[16] E. Ikkala, E. Hyvönen, H. Rantala, M. Koho, Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces, Semantic Web – Interoperability, Usability, Applicability 13 (2022) 69–84. doi:10.3233/SW-210428.

[17] S. Sarsa, E. Hyvönen, Searching case law judgements by using other judgements as a query, in: A. Filchenkov, J. Kauttonen, L. Pivovarova (Eds.), Artificial Intelligence and Natural Language. 9th Conference, AINL 2020, Helsinki, Finland, October 7–9, 2020, Springer-Verlag, 2020, pp. 145–157. URL: https://doi.org/10.1007/978-3-030-59082-6_11.

[18] R. Hoekstra, The MetaLex Document Server legal documents as versioned linked data, in: Proceedings of the ISWC 2011, Springer, 2011, pp. 128–143.

[19] N. Casellas, T. R. Bruce, S. S. Frug, S. Bouwman, D. Dias, J. Lin, S. Marathe, K. Rai, A. Singh, D. Sinha, S. Venkataraman, Linked legal data: Improving access to regulations, in: Proc. of the 13th Annual International Conf. on Digital Government Research (dg.o '12), Assoc. for Comp. Machinery, 2012, pp. 280–281.

[20] T. Burrows, N. B. Pinto, M. Cazals, A. Gaudin, H. Wijsman, Evaluating a Semantic Portal for the "Mapping Manuscript Migrations" Project, DigItalia 2 (2020) 178–185. URL: http://digitalia.sbn.it/article/view/2643.