Publishing, searching, and analyzing cross-border multilingual legislation on the Semantic Web

Eero Hyvönen[1,2][0000−0003−1695−5840], Hien Cao[1],
Rafael Leal[1][0000−0001−7266−2036], Heikki Rantala[1][0000−0002−4716−6564], and
Aki Hietanen[3]

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

Abstract. This paper concerns the problem of searching legislative documents in an international cross-broader multilingual setting. Here legal documents are published originally in different countries using different local languages, and the end-users are searching for the documents using their own languages. Furthermore, different country-specific semantic keyword and classification systems for indexing the contents may have been used. Cross-border services are needed, e.g., when moving from one country to another and looking for regulations for immigration, heath care, education, etc. To address the challenge, a cross-border solution based on Linked Open Data and Semantic Web technologies is presented, and a proof-of-concept system was designed and implemented, using consolidated laws of Finland and Estonia and EU directives as a case study. The demonstrator includes a semantic portal and a LOD service. Based on the so-called Sampo Model, the main novelty of the FinEstLawSampo demonstrator presented is the provision of heterogeneous cross-country, multilingual, distributed legal data through multiple application perspectives for faceted searching and exploring the data as well as for data analysis in legal informatics.

Keywords: Linked data, Law, Multilingual, Semantic portal, Data service

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 [25]. 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[4], and EU level systems, such as

[4] `https://www.legislation.gov.uk`

HUDOC[5], EUR Lex [6], the EU Cellar[7], and the ECLI Search Engine[8] for the case law.

In an international setting such as the EU, cross-border access to legislation published in different countries is often needed. Even though legislation is often available openly, it is not necessarily Findable, Accessible, Interoperable, and Reuseable according to the FAIR[9] for scientific data management and stewardship. A specific problem here is that local legislation and User Interfaces (UI) for searching and browsing it may be available only in local languages that the end-user does not understand. In addition, different local keyword vocabularies for subject matter indexing and classification systems are used in different countries, which sets challenges for querying the data semantically and for precision and recall of information retrieval. Furthermore, legal documents are often available only as texts for the humans to read with little semantic metadata available, which makes them hard to use in applications of legal informatics[10] [5], e.g., in computational law[11].

To address these problems, this paper argues that legislation should be published and used as Linked Open Data (LOD) on the Semantic Web, based on language-agnostic indexing schemes and/or by aligning local schemes onto each other. To address the problems of multilingualism, machine translation systems should be integrated in the search systems and UIs when human-made translations are not readily available. To support the argument, a model for publishing and using cross-border multilingual legislation databases is presented. As case study, integrating Finnish and Estonian legislation as well as EU Directives is considered by presenting a proof-of-concept system for searching, browsing, and studying law in an international setting. The data used is available as a LOD service and SPARQL endpoint[12] on top of which the portal was created[13].

In the following, we first describe our model for publishing and using cross-border legislation. After the LOD underlying FinEstLawSampo is presented and how it was created using, e.g., natural langugae processing (NLP). After this, the usage of the data service and the portal on top of it are explained. In conclusion, related works are discussed and contributions and lessons learned are summarized.

[5] https://hudoc.echr.coe.int/
[6] https://eur-lex.europa.eu/
[7] https://data.europa.eu/euodp/en/data/dataset/sparql-cellar-of-the-publications-office
[8] https://e-justice.europa.eu/content_ecli_search_engine-430-en.do
[9] https://www.go-fair.org/fair-principles/
[10] https://en.wikipedia.org/wiki/Legal_informatics
[11] https://law.stanford.edu/2021/03/10/what-is-computational-law/
[12] LOD data service online: https://ldf.fi/datasets/finestlaw
[13] Portal online: https://finestlaw.demo.seco.cs.aalto.fi/en

2 A Model for Publishing Cross-border Legislation

2.1 Sampo Model Applied

The project decided to test and use the Sampo model [12] and Sampo-UI framework [16,27] for designing and implementing FinEstLawSampo. The motivation from this came from the encouraging experiences of developing the LawSampo system for publishing Finnish legislation and case law on the semantic web [14]; the idea was to extend this existing application already in use with Estonian data for a new cross-border multilingual system.

The Sampo model consists of a set a general principles on 1) how to create LOD services and 2) user interfaces on top of them. The model has evolved gradually when developing over twenty LOD services and semantic portals[14] mostly in the domain of Cultural Heritage (CH) and DH.

Regarding LOD creation there are three major principles P1–P3 in the model.

1. *Support collaborative data creation and publishing (P1)* Leonardo da Vinci has said: *Learn how to see. Realize the everything connects to everything else.* This wisdom applies well to the general idea of LOD where mutually interlinked aggregated datasets are used to enrich each other, in our case Finnish and Estonian legislation and the underlying EU directives.
2. *Use a shared open ontology infrastructure (P2)* According to a wisdom of Albert Einstein *intellectual solve problems but geniuses prevent them.* This wisdom applies well to the idea of developing and using an infrastructure in creating CH and DH applications [13]: it is arguably better to prevent interoperability problems already when creating data than fix problems afterwards when aggregating data [10]. In our case, the LawSampo infrastructure could be re-used as well as EU level standards (e.g., ELI identifiers[15] [4]) and controlled vocabularies, especially the multilingual Pan-European EuroVoc thesaurus[16] maintained by the Publications Office of the European Union and hosted on the portal Europa.
3. *Make clear distinction between the LOD service and the user interface (UI) (P3)* This principle was tested first when developing the ontology service ONKI Light for SKOS vocabularies [30]: is it possible to re-implement the original ONKI.fi ontology services [34,37] by using SPARQL queries only for data access? The answer was "yes", and the Finto.fi ontology service was deployed based on ONKI Light. It was also tested whether it makes sense to apply this idea to implement faceted semantic search, too, used in Sampo systems since 2004. The answer was "yes" again, and this development lead to developing the tools SPAQRL Faceter [17] and later Sampo-UI in 2018 that has been used in some 15 Sampos.

As for the UI logic there are three principles P4–P6 in the Sampo model:

[14] Sampo systems homepage: `https://seco.cs.aalto.fi/applications/sampo/`

[15] `https://eur-lex.europa.eu/content/help/eurlex-content/eli.html`

[16] `https://en.wikipedia.org/wiki/EuroVoc`

1. *Provide multiple perspectives to the same data (P4)* The idea here is the same as in the FAIR principles[17], but adapted to UI design: reusing the data even within one UI. The class structure of KGs provide for this a natural approach: classes (e.g., Law, Directive, etc.) can be used as a basis for searching their individuals (particular laws, directives, etc) in the application perspectives.
2. *Standardize portal usage by a simple filter-analyze two-step cycle (P5)* This idea was inspired by the prosopograhical research method on groups of people [36], where a target group of people sharing some common features is first filtered out and then analyzed in more detail. This model is useful also for studying laws and directives.
3. *Support data analysis and knowledge discovery in addition to data exploration (P6)* In addition to semantic faceted search and data exploration, one should consider providing the user with intelligent tools for analyzing the data, or intelligent agents trying to find interesting pattern of knowledge in the data by themselves, solving research problems, and explaining the results to the user, leading to "third generation" systems in DH [11].

2.2 Federated Search Vs. Aggregating Global Data

When creating search and data exploration systems based on data aggregation, there are two basic approaches available.

1. *Distributed strategy*: federated search. The traditional way is to take the user's query, send it to distributed local data services hosting the data to be aggregated, collect the answers, and present them to the user.
2. *Centralized strategy*: aggregating global data. The other approach is to aggregate and harmonize the distributed heterogeneous datasets first into a global database or KG, and apply the query to its centralized data service.

In the distributed strategy, the burden of figuring out what the user wants can be distributed to the local data providers that transform the query for their local databases. Also the burden of actually executing the query can be distributed. Also centralized federated query processing is also used, like in SPARQL, but this can be computationally expensive. A challenge in federated search is that it is difficult to transform the query and present results in a semantically interoperable way in local services whose data models and vocabularies[18] used in the metadata are different. This deteriorates precision and recall, and makes data-analyses challenging. For example, entities, such as persons and places are typically represented in different ways locally and therefore confused with each other. Furthermore, not having simultaneous access to the global data is a severe restriction on what can be analyzed from the global data. For example, finding out relations between entities in local datasets is hard.

[17] `https://www.go-fair.org/fair-principles/`

[18] In this paper the term *vocabulary* is used to refer to (hierarchical) knowledge organization systems, such as thesauri, authority files, and geographical gazetteers, whose entries are used to fill in metadata element (property) values.

In the Sampo model used in FinEstLawSampo the centralized strategy was therefore selected, introduced already in the first Sampo system MuseumFinland [8] (online since 2004). However, using the global strategy brings in its own challenges. These regard especially data model harmonization of the local datasets and disambiguating and linking the data instances for semantic interoperability. However, these challenges are not due to the centralized strategy, but to the heterogeneity of the local datasets and the ways they are created, and have to be addressed in any case when dealing with local data in a semantically proper interoperable way.

2.3 Maintaining Linked Open Data and Data Services

According to Heraclitus (fl. 500 BC) *everything changes and nothing remains still; and you cannot step twice into the same stream.* An important issue of using LOD is maintaining changes in the KG as time goes by and software evolves. However, the Sampo principles discussed above focus only on how to create and publish a LOD service.

A piece of good news regarding the challenges of change is that linked data formats are open, standardized by W3C recommendations, and are based on text. The data is therefore pretty sustainable and re-usable, but tools, such as triple stores and UI frameworks change more often and may support and extend the standards, such as the SPARQL query language, in different ways. A more severe challenge is what to do, when either the metadata models [39], vocabularies used in populating the models, and the data itself evolves. This problem is discussed, e.g., in [18,19].

There are two basic approaches depending on how the primary data is managed. If the data is maintained in a legacy system using traditional formats, it makes sense to design the LOD transformation in such a way that it can be re-run automatically from scratch. This means that there should preferably be no intermediate manual phases in the process, as their results would be wiped away by when the KG is reconstructed. The challenge in this approach is that the new data is likely to contain typos and linking textual descriptions may need manual work and fixes after all. For finding quality issues, semantic validation languages and frameworks, such as SHACL[19] and ShEx[20] can be used.

A better way would be managing the KG in native linked data form. This would keep the data automatically consistent and ready to be uploaded into a LOD service. Unfortunately, there are still few tools for editing and managing RDF data. An exception to this are ontology editors, such as Protege[21] and Topbraid Composer[22]. In the case of the Sampo systems, the SPARQL SAHA editor [23] was developed and has been used in maintaining Sampo datasets by their data owners in some cases.

[19] `https://www.w3.org/TR/shacl/`
[20] `https://shex.io/`
[21] `https://protege.stanford.edu/`
[22] `https://allegrograph.com/topbraid-composer/`

3 Creating the Knowlegde Graph and LOD Service

This section overviews the data underlying FinEstLawSampo how it was transformed into a KG, and publihised as a LO service.

3.1 Primary Data and Data Model

Finnish legislation and case law decisions have been published as web documents since 1997 in the Finlex Data Bank[23]. 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 [24] LOD service that currently contains ca. 28 million triples. In LawSampo, 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.

The main classes of the simple data model are shown in Table 1 with the number of instances and descriptions for each class. 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[24]. The data model schema is available and documented at the namespace URI `http://ldf.fi/schema/lawsampo/`.

Table 1: The main classes of FinEstLawSampo

Class	**Description**
:Statute	A statute in consolidated legislation
:Section	A section of a statute in consolidated legislation
:SituationCategory	A life situation category of a document
:EuroVocKeyword	A EuroVoc keyword of a document

In FinEstLawSampo this data model was reused also for the Estonian statutes that were available in custom XML format[25], in Estonian and in English. The XML documents were transformed into the LawSampo RDF data model using Python SAX APIs [26] and Python RDFLib [27].

The final KG contains 12 394 Finnish statutes and 351 Estonian ones. In addition, metadata about 4972 EU directives from the EU Cellar were imported the system.

[23] `http://www.finlex.fi`

[24] `https://www.dublincore.org/specifications/dublin-core/dcmi-terms/`

[25] Consolidated texts of of Estonian legislation: `https://www.riigiteataja.ee/en/`

[26] `https://docs.python.org/3/library/xml.sax.handler.html`

[27] `https://github.com/RDFLib/rdflib`

3.2 Data Enrichment

The original datasets were enriched by data linking and by natural language processing:

Internal and External 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.

Both Finnish and Estonian statutes were linked externally to EU directives mentioned in them available at the EU Cellar[28]. Cellar is the common data repository of the Publications Office of the European Union. Cellar stores multilingual publications and metadata, is open to all EU citizens, and provides both human- and machine-readable data in RDF form.

In order to further process the legal documents in this project, four main Natural Language Processing (NLP) techniques were used: 1) their contents were translated automatically to cover all wanted languages (Finnish, Estonian, English); 2) keywords were extracted; 3) the documents were classified using two different sets of Life Events and their similarity was calculated.

Translations The objective of the portal is to offer the same content in three different languages: Finnish, Estonian, and English. This means that translations have to be provided for all desired language pairings (from Finnish to Estonian and English, and from Estonian to Finnish and English). Out of these, the original data contains official translations from Estonian to English, but all other pairings are missing. We decided to use automatic machine translation to fill this gap, even though these translations are not of legal tender.

The translations were carried out using Opus-MT's machine translation models [32] built by the Language Technology Research Group of the University of Helsinki. These are a series of deep learning translation models, each one finetuned for a specific language pair. We used the corresponding models [29] via Huggingface's Transformers library [38].

These models typically have a maximum input length of 512 tokens, which are roughly equivalent to 400 words in English or 300 in Finnish. This means that only very short texts can be translated at once, which would be the ideal situation since all the context could be preserved. In order to overcome this problem but still provide some context to the translations, the documents were separated into sentences and the previous sentence was fed to the model alongside the sentence to be translated, except when starting a new paragraph. Heuristics were applied in case the model behaved unexpectedly. Some hallucination is still present in the translations of overlong sentences.

[28] `https://op.europa.eu/en/web/cellar`

[29] `Helsinki-NLP/opus-mt-tc-big-fi-en`, `Helsinki-NLP/opus-mt-fi-et` and `Helsinki-NLP/opus-mt-et-fi`

In order to preserve the original HTML format in the translations, the Python library XML2Dict [30] was used to unpack and repack the sentences.

Keyword extraction The objective of extracting keywords is twofold: on the one hand, it allows the users of the portal to navigate documents succinctly and conveniently by choosing keywords of interest. On the other hand, it provides an internal representation for the documents, which allows for both classification among Life Events and text similarity assessment.

EuroVoc[31] was chosen as the keyword vocabulary due to its international, cross-border nature and its strong support for legal texts, given its purpose to cover EU institutions and activities. It contains keyword labels in 24 EU languages, which makes translating them trivial. The keywords are extracted using the third-party tool *PyEuroVoc* [1].

Classification Based on Life Situations A novelty of FinEstLawSampo is to provide the end user the statues from a perspective of life situations the end user is expected be involved. Both in Finland[32] and in Estonia[33], public administrations are using such classifications in order to provide their services is an natural user friendly way. For example, the Finnish system includes the nine major situations below, with refined sub-situations):

1. *Living together and having a family*: Living together; Having children; Welcome to adulthood!; Divorce or separation; Death of a close family member
2. *Social security*: Guardianship; Informal carer for a loved one; Retirement; Services for people with disabilities; Services for the elderly; Income support
3. *Health and medical care*: (Staying healthy; Falling ill; Nutrition and food; Rehabilitation; Substance abuse and gambling; Coronavirus
4. *Teaching and education*: Pre-primary education and schooling; Studying; Livelihood and social assistance of students; Science and research
5. *Working life and unemployment*: Unemployment; Starting a business; Rules of working life
6. *Housing and construction*: Purchasing a home; Construction and properties;
7. *Rights and obligations*: Fundamental rights and civic activity; Legislation and legal protection; Court proceedings and criminal matters; Security and public order; Digital support and administrative services; Data leak
8. *Personal finances*: Managing your personal finances; Taxation and public finances; Consumer protection
9. *Moving and travelling*: Work in Finland; Migration; Travel

[30] `https://github.com/mcspring/XML2Dict`

[31] `https://data.europa.eu/data/datasets/eurovoc`

[32] Finnish classification of life events: `https://www.suomi.fi/citizen`

[33] Estonian classification of life events: `https://www.eesti.ee/en` (on the left sidebar, under the header "Citizen")

The Estonian counterpart has a different set of 12 life event categories. These may list individual laws, but not in any case exhaustively.

In the FinEstLawSampo KG all legal documents were automatically classified into both Finnish and Estonian life events categories. As a result, all Finnish legal documents are classified using the set of categories in the Finnish portal as well as the Estonian portal, and vice-versa. These categorizations are not exclusive but multi-label: a document may belong to different categories within to the same set.

In order to carry out the classification, the keywords are used as the basis for the internal text representation of the system, as described in more detail in [20]. The keywords are translated to Finnish where needed and transformed into word embeddings via the corresponding pre-trained fastText model [22], which are then pooled together to create the document representation. Translating to Finnish is important in order to allow for semantic reinforcement of the categories using the General Finnish Ontogoly YSO[34] [20] as well as for measuring text similarity, which is a simple cosine similarity calculation.

Linked Open Data Service The FinEstLawSampo data service adopts the 5-star Linked Data model[35], extended with two more stars, as suggested in the Linked Data Finland model and platform [15]. The 6th star is obtained by providing the dataset schemas and documenting them. The FinEstLawSampo schema can be downloaded from the service[36] and the data model is documented using the LODE service[37]. The 7th star is achieved by validating the data against the documented schemas to prevent errors in the published data. FinEstLawSampo attempts to obtain the 7th star by applying different means of combing out errors in the data within the data conversion process. The FinEstLawSampo data model and its integrity constraints are presented in a machine-processable format using the ShEx Shape Expressions language[38] [31]. We have made initial validation experiments with the PyShEx[39] 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[40] 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[41]. As the triplestore, Apache

[34] `https://finto.fi/yso/en/`

[35] `https://www.w3.org/DesignIssues/LinkedData.html`

[36] `https://www.ldf.fi/dataset/lawsampo`

[37] `https://essepuntato.it/lode/`

[38] `https://shex.io`

[39] `https://github.com/hsolbrig/PyShEx`

[40] `http://ldf.fi`

[41] `https://ldf.fi/lawsampo/sparql`

Jena Fuseki[42] is used as a Docker container, which allows efficient provisioning of resources (CPU, memory), portability, and scaling. Varnish Cache web application accelerator[43] is used for routing URIs, content negotiation, and caching.

4 Using the FinEstLawSampo Portal and LOD Service

After a Sampo LOD service has been established it can be used in two ways:

1. *Using Application Programming Interfaces (API).* The LOD publication methodology provides different ways to access the data: 1) The data can be downloaded from the service as data dumps. 2) The data can be browsed in a human readable way using a linked data browser[44]. 3) The LOD service provides content negotiation where URIs can be resolved and either data for the machine or HTML for the human user can be returned[45]. 4) Most importantly, the data service can be queried in flexible ways using the SPARQL query language[46] and endpoint. There are easy to use tools, such as Yasgui [28], for editing end executing SPARQL queries with some built-in visualization options for the results. The SPARQL endpoint can be accessed from any programming environment, such as Jupyter notebooks and Python scripting for querying and analyzing data.
2. *Using portals and other applications.* Ready-to-use applications for accessing and using the data without programming skills can be developed on top of the LOD service, as exemplified by the Sampo portal series.

In a Sampo-UI-based portal the user first lands on the *landing page* with several *application perspectives* to the data. The perspectives are based on classes of the underlying KG, in our case statutes and EU directives. The usage cycle of each perspective can be divided into two steps: 1) filter and 2) analyze. The user first filters the data by using the faceted semantic search [6,33] tools provided by the portal. The results as well as the facet option hit counts are updated after each category selection on a facet. In faceted search, the hit counts direct the search and prevent ending up in dead-end situations where no results are found. Faceted search was developed already in the 90's and early 00's but under the name "view-based search" [26,9] and also as "dynamic taxonomies" [29].

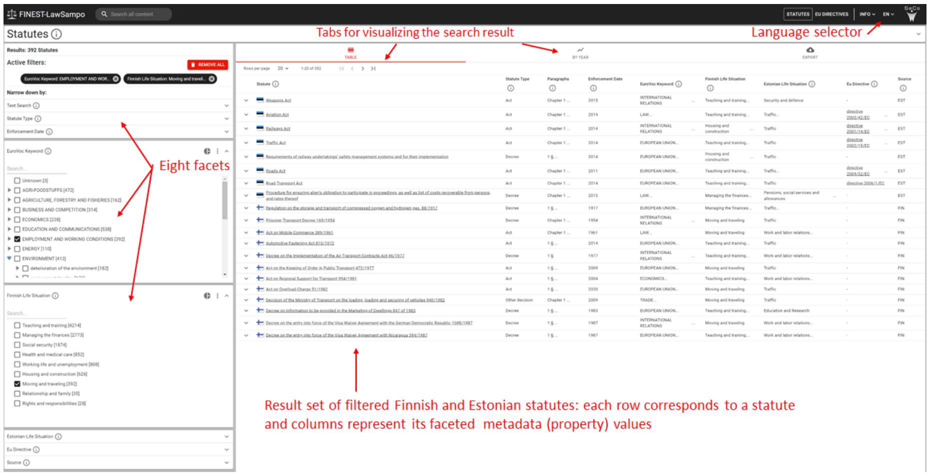
After filtering the data to the wanted subset, the *target group*, the user can analyze the results set, i.e., a set of instances of the class corresponding to the application perspective, with integrated data-analytic tools available as tabs on the application perspective page.

It is also possible to select a particular instance of the result set for a closer look: each instance has an *instance page* that provides aggregated information

[42] `https://jena.apache.org/documentation/fuseki2/`
[43] `https://varnish-cache.org`
[44] See, e.g., the browser for DBpedia: `https://dbpedia.org/ontology/Browser`
[45] Content Negotiation by Profile: `https://www.w3.org/TR/dx-prof-conneg/`
[46] SPARQL 1.1 Query Language: `https://www.w3.org/TR/sparql11-query/`

about the individual with internal and external links for further information to browse. Instance pages also may have a set of tabs that provide contextualized data-analyses of the individuals in the same way as for target groups.

This filter-analyze two-step usage cycle allows an iterative approach to exploring the data [21,35]. It is possible to find potentially interesting subsets and individuals in the data without having to be already familiar with the content. By providing a text facet, it is also possible to support use cases where the user is looking for a specific instance, say a person with a known name, and can formulate the search query easily.



Tabs for visualizing the search result

Language selector

Eight facets

Results: 392 Statutes

Active filters: EuroVoc Keyword: EMPLOYMENT AND WOR... Finnish Life Situation: Moving and traveli...

Narrow down by:

Text Search

Statute Type

Enforcement Date

EuroVoc Keyword

- □ Unknown [3]
- □ AGRI-FOODSTUFFS [472]
- □ AGRICULTURE, FORESTRY AND FISHERIES [162]
- □ BUSINESS AND COMPETITION [314]
- □ ECONOMICS [228]
- □ EDUCATION AND COMMUNICATIONS [538]
- ☑ EMPLOYMENT AND WORKING CONDITIONS [392]
- □ ENERGY [110]
- □ ENVIRONMENT [412]
  - □ deterioration of the environment [182]

Finnish Life Situation

- □ Teaching and training [4214]
- □ Managing the finances [2773]
- □ Social security [1874]
- □ Health and medical care [852]
- □ Working life and unemployment [808]
- □ Housing and construction [626]
- ☑ Moving and traveling [392]
- □ Relationship and family [35]
- □ Rights and responsibilities [28]

Estonian Life Situation

Eu Directive

Source

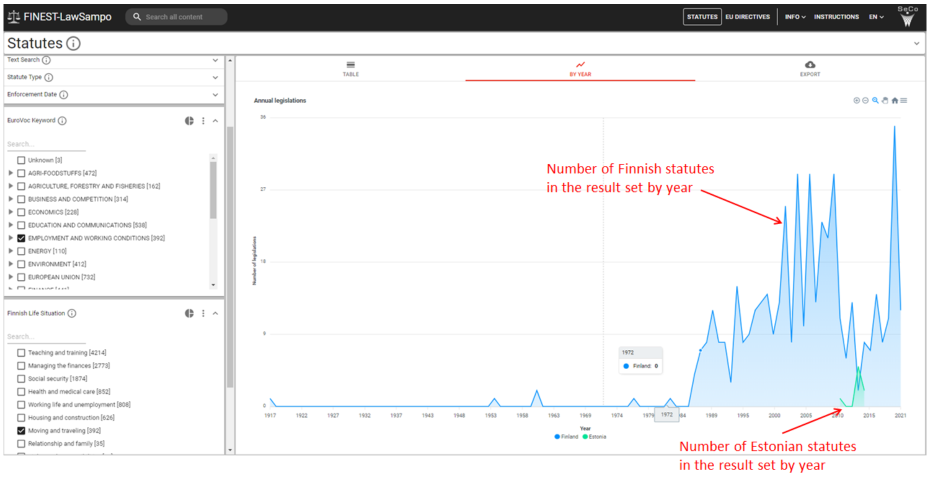
Statute	Statute Type	Paragraphs	Enforcement Date	EuroVoc Keyword	Finnish Life Situation	Estonian Life Situation	Eu Directive	Source
Weapons Act	Act	Chapter 1 ...	2015	INTERNATIONAL RELATIONS	Teaching and training...	Security and defence	-	EST
Aviation Act	Act	Chapter 1 ...	2014	LAW...	Teaching and training...	Traffic...	directive 2003/42/EC	EST
Railways Act	Act	Chapter 1 ...	2014	INTERNATIONAL RELATIONS	Housing and construction	Traffic	directive 2001/14/EC	EST
Traffic Act	Act	Chapter 1 ...	2014	EUROPEAN UNION...	Teaching and training...	Traffic	directive 2002/15/EC	EST
Requirements of railway undertakings' safety management systems and for their implementation	Decree	1 § ...	2014	EUROPEAN UNION...	Housing and construction	Traffic	-	EST
Roads Act	Act	Chapter 1 ...	2011	EUROPEAN UNION...	Teaching and training...	Traffic...	directive 2004/52/EC	EST
Road Transport Act	Act	Chapter 1 ...	2014	EUROPEAN UNION...	Teaching and training...	Traffic	directive 2006/1/EC	EST
Procedure for ensuring alien's obligation to participate in proceedings, as well as list of costs recoverable from persons, and rates thereof	Decree	Chapter 1 ...	2015	LAW...	Managing the finances...	Pensions, social services and allowances	-	EST
Regulation on the storage and transport of compressed oxygen and hydrogen gas. 88/1917	Decree	1 § ...	1917	EUROPEAN UNION...	Managing the finances...	Traffic...	-	FIN
Prisoner Transport Decree 165/1954	Decree	Chapter 1 ...	1954	INTERNATIONAL RELATIONS	Moving and traveling	Traffic	-	FIN
Act on Mobile Commerce 389/1961	Act	Chapter 1 ...	1961	LAW...	Moving and traveling	Work and labor relations...	-	FIN
Automotive Fastening Act 810/1972	Act	1 § ...	2014	EUROPEAN UNION...	Teaching and training...	Traffic	-	FIN
Decree on the Implementation of the Air Transport Contracts Act 46/1977	Decree	1 §	1977	INTERNATIONAL RELATIONS	Teaching and training...	Work and labor relations...	-	FIN
Act on the Keeping of Order in Public Transport 472/1977	Act	1 § ...	2009	EUROPEAN UNION...	Moving and traveling	Traffic	-	FIN
Act on Regional Support for Transport 954/1981	Act	1 § ...	2004	ECONOMICS...	Teaching and training...	Work and labor relations...	-	FIN
Act on Overload Charge 51/1982	Act	1 § ...	2020	EUROPEAN UNION...	Moving and traveling	Traffic	-	FIN
Decision of the Ministry of Transport on the loading, loading and securing of vehicles 940/1982	Other decision	Chapter 1 ...	2009	TRADE...	Moving and traveling	Traffic	-	FIN
Decree on Information to be provided in the Marketing of Dwellings 847 of 1983	Decree	1 § ...	1983	EUROPEAN UNION...	Teaching and training...	Education and Research	-	FIN
Decree on the entry into force of the Visa Waiver Agreement with the German Democratic Republic 1088/1987	Decree	1 § ...	1987	INTERNATIONAL RELATIONS	Moving and traveling	Work and labor relations...	-	FIN
Decree on the entry into force of the Visa Waiver Agreement with Nicaragua 394/1987	Decree	1 § ...	1987	EUROPEAN UNION...	Teaching and training...	Work and labor relations...	-	FIN

Result set of filtered Finnish and Estonian statutes: each row corresponds to a statute and columns represent its faceted metadata (property) values

Fig. 1: Faceted search for Finnish and Estonian statutes

The landing page of the FinEstLawSampo portal offers an application perspective for 12 745 Finnish and Estonian status and an another similar one for 4972 EU directives. By clicking on the Statutes perspective box on the landing page, a faceted search interface for searching and browsing statutes is opened (Fig. 1). The eight facets on the left are based on the property values of the class Statute and include: 1) Traditional text search facet, 2) Statute type, 3) Enforcement date, 4) EuroVoc keyword, 5) Finnish life situation, 6) Estonian life situation, 7) EU directive (mentioned), and 8) Source (of legislation). The user has selected from the EuroVoc keyword facet EMPLOYMENT AND WORKING CONDITIONS and from the Finnish life situation facet Moving and atutes found are show on TABLE tab on the left with country flags.

As customary in the Sampo-UI model, the search results can be analyzed on different tabs of the application perspective. By selecting the tab BY YEAR the result set could be visualized on a timeline based on the enforcement date of the statutes. This gives the user contextual information on how the statutes in the search result set of interest has evolved in time (cf. Fig. 2).



Text Search
Statute Type
Enforcement Date

EuroVoc Keyword

Search...

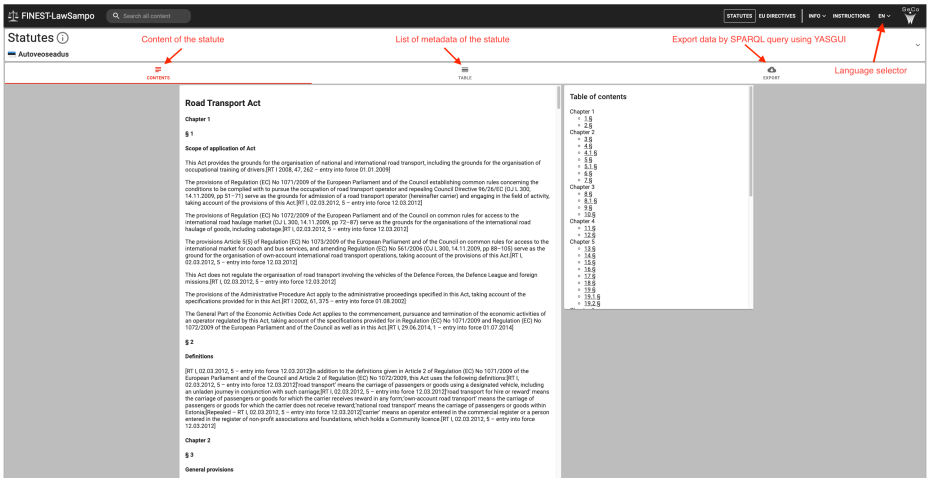
- □ Unknown [3]
- □ AGRI-FOODSTUFFS [472]
- □ AGRICULTURE, FORESTRY AND FISHERIES [162]
- □ BUSINESS AND COMPETITION [314]
- □ ECONOMICS [228]
- □ EDUCATION AND COMMUNICATIONS [538]
- ☑ EMPLOYMENT AND WORKING CONDITIONS [392]
- □ ENERGY [110]
- □ ENVIRONMENT [412]
- □ EUROPEAN UNION [732]

Finnish Life Situation

Search...

- □ Teaching and training [4214]
- □ Managing the finances [2773]
- □ Social security [1874]
- □ Health and medical care [852]
- □ Working life and unemployment [808]
- □ Housing and construction [626]
- ☑ Moving and traveling [392]
- □ Relationship and family [35]

Fig. 2: Visualizing the number of statutes in the result set on a timeline by their enforcement day



Road Transport Act

Chapter 1

§ 1

Scope of application of Act

This Act provides the grounds for the organisation of national and international road transport, including the grounds for the organisation of occupational training of drivers.[RT I 2008, 47, 262 – entry into force 01.01.2009]

The provisions of Regulation (EC) No 1071/2009 of the European Parliament and of the Council establishing common rules concerning the conditions to be complied with to pursue the occupation of road transport operator and repealing Council Directive 96/26/EC (OJ L 300, 14.11.2009, pp 51–71) serve as the grounds for admission of a road transport operator (hereinafter carrier) and engaging in the field of activity, taking account of the provisions of this Act.[RT I, 02.03.2012, 5 – entry into force 12.03.2012]

The provisions of Regulation (EC) No 1072/2009 of the European Parliament and of the Council on common rules for access to the international road haulage market (OJ L 300, 14.11.2009, pp 72–87) serve as the grounds for the organisations of the international road haulage of goods, including cabotage.[RT I, 02.03.2012, 5 – entry into force 12.03.2012]

The provisions Article 5(5) of Regulation (EC) No 1073/2009 of the European Parliament and of the Council on common rules for access to the international market for coach and bus services, and amending Regulation (EC) No 561/2006 (OJ L 300, 14.11.2009, pp 88–105) serve as the ground for the organisation of own-account international road transport operations, taking account of the provisions of this Act.[RT I, 02.03.2012, 5 – entry into force 12.03.2012]

This Act does not regulate the organisation of road transport involving the vehicles of the Defence Forces, the Defence League and foreign missions.[RT I, 02.03.2012, 5 – entry into force 12.03.2012]

The provisions of the Administrative Procedure Act apply to the administrative proceedings specified in this Act, taking account of the specifications provided for in this Act.[RT I 2002, 61, 375 – entry into force 01.08.2002]

The General Part of the Economic Activities Code Act applies to the commencement, pursuance and termination of the economic activities of an operator regulated by this Act, taking account of the specifications provided for in Regulation (EC) No 1071/2009 and Regulation (EC) No 1072/2009 of the European Parliament and of the Council as well as in this Act.[RT I, 29.06.2014, 1 – entry into force 01.07.2014]

§ 2

Definitions

[RT I, 02.03.2012, 5 – entry into force 12.03.2012]In addition to the definitions given in Article 2 of Regulation (EC) No 1071/2009 of the European Parliament and of the Council and Article 2 of Regulation (EC) No 1072/2009, this Act uses the following definitions:[RT I, 02.03.2012, 5 – entry into force 12.03.2012]'road transport' means the carriage of passengers or goods using a designated vehicle, including an unladen journey in conjunction with such carriage;[RT I, 02.03.2012, 5 – entry into force 12.03.2012]'road transport for hire or reward' means the carriage of passengers or goods for which the carrier receives reward in any form;'own-account road transport' means the carriage of passengers or goods for which the carrier does not receive reward;'national road transport' means the carriage of passengers or goods within Estonia;[Repealed – RT I, 02.03.2012, 5 – entry into force 12.03.2012]'carrier' means an operator entered in the commercial register or a person entered in the register of non-profit associations and foundations, which holds a Community licence.[RT I, 02.03.2012, 5 – entry into force 12.03.2012]

Chapter 2

§ 3

General provisions

Table of contents

Chapter 1
- 1 §
- 2 §

Chapter 2
- 3 §
- 4 §
- 4.1 §
- 5 §
- 5.1 §
- 6 §
- 7 §

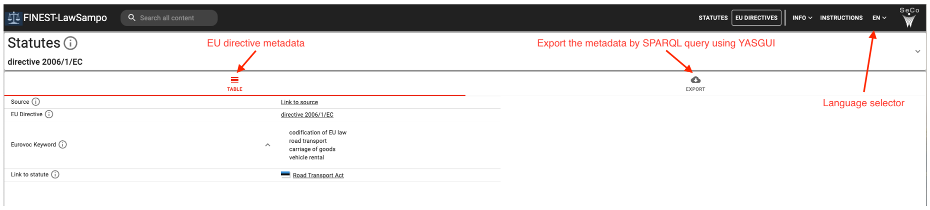
Chapter 3
- 8 §
- 8.1 §
- 9 §
- 10 §

Chapter 4
- 11 §
- 12 §

Chapter 5
- 13 §
- 14 §
- 15 §
- 16 §
- 17 §
- 18 §
- 19 §
- 19.1 §
- 19.2 §

Fig. 3: Homepage of a single statute with tabs

By clicking on a statute link its home page is opened for close reading (Fig. 3). This page shows the statute in detail and its metadata including 1) statute (name of the legislation), 2) statute type, paragraphs (table of content of the legislation), 3) enforcement date, 4) EuroVoc keyword, 5) Finnish life situation, 6) Estonian life situation, 7) EU directive (link to EU directive metadata view), 8) similar statutes (links to similar legislation in the other country), 9) Source (of legislation), 10) Link to original source (of legislation). English, Finnish or Estonian language can be selected as the interface language at the top right corner, and the facets and statute texts are automatically translated if needed into the language of preference. A disclaimer in shown on the page if the content was machine-translated. On the tab EXPORT, a window for the Yasqui SPARQL editor is opened with a query for retrieving the statute in question. This feature, available also at the main application perspective page (Fig. 1), is provided for users interested in learning how to qyery the data by themselves.



Statutes

directive 2006/1/EC

Source	Link to source
EU Directive	directive 2006/1/EC
Eurovoc Keyword	codification of EU law road transport carriage of goods vehicle rental
Link to statute	Road Transport Act

Fig. 4: Homepage of a single EU directive, including a link to full source data in EU Cellar

On the landing page, there is also another application perspective for finding the EU directives with faceted search. This application can be accessed by clicking the EU directive perspective box. There is a facet for EuroVoc keywords that can be used to filter EU directives based on their subject matter content. By clicking a EUn directive in the result table, a homepage with metadata about the EU directive is shown (Fig. 4). The EU directive metadata includes 1) Source (to the original content of the EU directive), 2) EuroVoc keywords, 3) Links to the Finnish and Estonian statutes mention EU directive in their textual contents.

5 Discussion

Related Works Our work on legal Linked Data services was influenced by the MetaLex Document Server[47] [7] and related national online services for legal documents in Greece, Luxemburg[48], France, Norway[49], and the U.S. [3]. 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[50] in the U.S.

A major point of reference of our work is the cross-border N-Lex system[51]. This multilingual system is based on federated search using various local national

[47] http://doc.metalex.eu

[48] http://legilux.public.lu/editorial/eli

[49] http://lovdata.no/eli

[50] https://scholar.google.com/scholar_courts

[51] https://n-lex.europa.eu/n-lex/

legal web services. Automatic translation functions between different languages are provided. When using this system, query words are translated into local languages or EuroVoc-based translations can be used, the search is the performed at local services, and the results can also be translated automatically if needed. The challenge of a system like this is that already the query word translation is challenging due to ambiguities not to mention text search at local web services based on different data models, vocabularies, and languages. FinEstLawSampo is a approach to address these issues by using Linked Data and the global data aggregation strategy.

Contributions, Evaluation, 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. FinEstLawSampo has access only to the latest versions of manually consolidated statues available in Finlex, and only retrieves Estonian legislation database in Riigiteataja[52] in a specific date. The problem of finding out how the statutes may have changed in time is left to the end user.

Usability of the FinEstLawSampo Portal has not been evaluated yet. However, the Sampo model has been evaluated in some other Sampo portals [2] 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 [12].

An informal evaluation of the portal suggests that relevant legislation can be found without lots of garbage, but a more accurate evaluation is challenging. There is no gold standard available and determining precision and recall of search would require substantial legal international expertise. It should be noted that

[52] https://www.riigiteataja.ee/avaandmed/ERT/

the portal also includes a traditional text search facet an option to use, and from that perspective it is more versatile that pure text search systems. In a system like FinEstLawSampo, good recall provided by the other alternative facets is probably more important than precision, as the end user is probably interested in exploring the legislation data (s)he is not familiar with.

In spite of the challenges and complexities of the underlying data and the use case, we believe that that proposed LOD approach is feasible and usable in practice, although building systems like FinEstLawSampo is more demanding than simple federated search systems and requires more collaboration and agreements and standards between the national legal data publishers.

Acknowledgments This research was part of the Nordic-Baltic project Achieving the World's Smoothest Cross-Border Mobility and Daily Life Through Digitalisation, funded by the Nordic Council of Ministers and the Cross-border Digital Services (CDBS) programme. Thanks to Anne Kari of the Finnish Digital Agency for collaborations. CSC – IT Center for Science provided computational resources for the project.

References

1. Avram, A.M., Pais, V., Tufis, D.I.: PyEuroVoc: A Tool for Multilingual Legal Document Classification with EuroVoc Descriptors. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). pp. 92–101. INCOMA Ltd., https://aclanthology.org/2021.ranlp-1.12
2. Burrows, T., Pinto, N.B., Cazals, M., Gaudin, A., Wijsman, H.: Evaluating a Semantic Portal for the "Mapping Manuscript Migrations" Project. DigItalia **2**, 178–185 (2020), http://digitalia.sbn.it/article/view/2643
3. Casellas, N., Bruce, T.R., Frug, S.S., Bouwman, S., Dias, D., Lin, J., Marathe, S., Rai, K., Singh, A., Sinha, D., Venkataraman, S.: Linked legal data: Improving access to regulations. In: Proc. of the 13th Annual International Conf. on Digital Government Research (dg.o '12). pp. 280–281. Assoc. for Comp. Machinery (2012)
4. Council of the European Union: Council conclusions inviting the introduction of the European Legislation Identifier (ELI). In: Official Journal of the European Union, C 325, 26.10.2012. pp. 3–11. Publications Office of the EU (2012)
5. Erdelez, S., O'Hare, S.: Legal informatics: Application of information technology in law. Annual Review of Information Science and Technology **32** (01 1997)
6. Hearst, M.: Design recommendations for hierarchical faceted search interfaces. In: ACM SIGIR workshop on faceted search. pp. 1–5. Seattle, WA (2006)
7. Hoekstra, R.: The MetaLex Document Server legal documents as versioned linked data. In: Proceedings of the ISWC 2011. pp. 128–143. Springer (2011)
8. Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M., Kettula, S.: MuseumFinland—Finnish museums on the semantic web. Journal of Web Semantics **3**(2), 224–241 (2005). https://doi.org/10.1016/j.websem.2005.05.008
9. Hyvönen, E., Saarela, S., Viljanen, K.: Application of ontology techniques to view-based semantic search and browsing. In: The Semantic Web: Research and Applications. Proceedings of the First European Semantic Web Symposium (ESWS 2004). Springer (2004). https://doi.org/10.1007/978-3-540-25956-5_7

10. Hyvönen, E.: Preventing interoperability problems instead of solving them. Semantic Web – Interoperability, Usability, Applicability **1**(1–2), 33–37 (2010). https://doi.org/10.3233/SW-2010-0014
11. Hyvönen, E.: Using the Semantic Web in Digital Humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. Semantic Web – Interoperability, Usability, Applicability **11**(1), 187–193 (2020). https://doi.org/10.3233/SW-190386
12. Hyvönen, E.: Digital humanities on the Semantic Web: Sampo model and portal series. Semantic Web – Interoperability, Usability, Applicability pp. 1–16 (2022), https://doi.org/10.3233/SW-223034
13. Hyvönen, E.: How to create a national cross-domain ontology and linked data infrastructure and use it on the semantic web. Semantic Web – Interoperability, Usability, Applicability (2023), https://seco.cs.aalto.fi/publications/2022/hyvonen-infra-2022.pdf, under review
14. Hyvönen, E., Tamper, M., Oksanen, A., Ikkala, E., Sarsa, S., Tuominen, J., Hietanen, A.: 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. pp. 110–114. Springer (2019)
15. Hyvönen, E., Tuominen, J., Alonen, M., Mäkelä, E.: Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets. In: ESWC 2014 Satellite Events. pp. 226–230. Springer (2014). https://doi.org/10.1007/978-3-319-11955-7_24
16. Ikkala, E., Hyvönen, E., Rantala, H., Koho, M.: Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces. Semantic Web – Interoperability, Usability, Applicability **13**(1), 69–84 (January 2022). https://doi.org/10.3233/SW-210428
17. Koho, M., Heino, E., Hyvönen, E.: SPARQL Faceter—Client-side Faceted Search Based on SPARQL. In: Troncy, R., Verborgh, R., Nixon, L., Kurz, T., Schlegel, K., Vander Sande, M. (eds.) Joint Proc. of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop. CEUR Workshop Proceedings, Vol-1615 (2016), http://ceur-ws.org/Vol-1615/semdevPaper5.pdf
18. Koho, M., Ikkala, E., Heino, E., Hyvönen, E.: Maintaining a linked data cloud and data service for second world war history. In: Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection. EuroMed 2018. Lecture Notes in Computer Science, vol. 11196. Springer (2018). https://doi.org/10.1007/978-3-030-01762-0_12
19. Koho, M., Ikkala, E., Hyvönen, E.: How to maintain a linked data cloud in a deployed semantic portal. In: Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks. CEUR Workshop Proceedings, Vol. 2180 (2018), http://ceur-ws.org/Vol-2180/
20. Leal, R., Kesäniemi, J., Koho, M., Hyvönen, E.: Relevance Feedback Search Based on Automatic Annotation and Classification of Texts. In: 3rd Conference on Language, Data and Knowledge (LDK 2021). Open Access Series in Informatics (OASIcs), vol. 93, pp. 18:1–18:15. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany (2021). https://doi.org/10.4230/OASIcs.LDK.2021.18
21. Marchionini, G.: Exploratory search: from finding to understanding. Communications of the ACM **49**(4), 41–46 (2006). https://doi.org/10.1145/1121949.1121979
22. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A.: Advances in pre-training distributed word representations. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018) (2018)

23. Mäkelä, E., Hyvönen, E.: SPARQL SAHA, a configurable linked data editor and browser as a service. In: Proceedings of the ESWC 2014 demonstration track, Springer-Verlag (2014)
24. Oksanen, A., Tuominen, J., Mäkelä, E., Tamper, M., Hietanen, A., Hyvönen, E.: 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, pp. 212–228. IOS Press (2019)
25. van Opijnen, M., Peruginelli, G., Kefali, E., Palmirani, M.: Online publication of court decisions in europe. Legal Information Management **17**, 136—-145 (2017). https://doi.org/10.1017/S1472669617000299
26. Pollitt, A.S.: The key role of classification and indexing in view-based searching. Tech. rep., University of Huddersfield, UK (1998), http://www.ifla.org/IV/ifla63/63polst.pdf
27. Rantala, H., Ahola, A., Ikkala, E., Hyvönen, E.: How to create easily a data analytic semantic portal on top of a SPARQL endpoint: introducing the configurable Sampo-UI framework. In: VOILA! 2023 Visualization and Interaction for Ontologies, Linked Data and Knowledge Graphs 2023. CEUR Workshop Proceedings, Vol. 3508 (2023), https://ceur-ws.org/Vol-3508/paper3.pdf
28. Rietveld, L., Hoekstra, R.: The YASGUI family of SPARQL clients. Semantic Web – Interoperability, Usability, Applicability **8**(3), 373–383 (2017)
29. Sacco, G.M.: Dynamic taxonomies: guided interactive diagnostic assistance. In: Wickramasinghe, N. (ed.) Encyclopedia of Healthcare Information Systems. Idea Group (2005)
30. Suominen, O., Johansson, A., Ylikotila, H., Tuominen, J., Hyvönen, E.: Vocabulary services based on SPARQL endpoints: ONKI Light on SPARQL. In: Poster proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2012) (2012), https://seco.cs.aalto.fi/publications/2012/suominen-et-al-onkilight-2012.pdf
31. Thornton, K., Solbrig, H., Stupp, G.S., Gayo, J.E.L., Mietchen, D., Prud'hommeaux, E., Waagmeester, A.: Using shape expressions (ShEx) to share RDF data models and to guide curation with rigorous validation. In: The Semantic Web. ESWC 2019. pp. 606–620. Springer (2019). https://doi.org/10.1007/978-3-030-21348-0_39
32. Tiedemann, J., Thottingal, S.: OPUS-MT – Building open translation services for the World. In: Martins, A., Moniz, H., Fumega, S., Martins, B., Batista, F., Coheur, L., Parra, C., Trancoso, I., Turchi, M., Bisazza, A., Moorkens, J., Guerberof, A., Nurminen, M., Marg, L., Forcada, M.L. (eds.) Proceedings of the 22nd Annual Conference of the European Association for Machine Translation. pp. 479–480. European Association for Machine Translation (2020), https://aclanthology.org/2020.eamt-1.61
33. Tunkelang, D.: Faceted Search. Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool, Palo Alto, CA, USA (2009). https://doi.org/10.1007/978-3-031-02262-3
34. Tuominen, J., Frosterus, M., Viljanen, K., Hyvönen, E.: ONKI SKOS server for publishing and utilizing SKOS vocabularies and ontologies as services. In: Proceedings of the 6th European Semantic Web Conference (ESWC 2009). Springer (2009)
35. Tzitzikas, Y., Manolis, N., Papadakos, P.: Faceted exploration of rdf/s datasets: a survey. Journal of Intelligent Information Systems **48**(2), 329–364 (2017)

36. Verboven, K., Carlier, M., Dumolyn, J.: A short manual to the art of prosopography. In: Prosopography approaches and applications. A handbook, pp. 35–70. Unit for Prosopographical Research (Linacre College) (2007). https://doi.org/1854/8212
37. Viljanen, K., Tuominen, J., Hyvönen, E.: Ontology libraries for production use: The Finnish ontology library service ONKI. In: Proceedings of the ESWC 2009, Heraklion, Greece. pp. 781–795. Springer (2009)
38. Wolf, T., Debut, L., Sanh, V., Chaumond, J., et al.: Transformers: State-of-the-Art Natural Language Processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6
39. Zeng, M., Qin, J.: Metadata, Third Edition. ALA Neal-Schuman, Chicago (2022)