A Model and Case Study for Searching and Reading Cross-border Multilingual Legislation on the Semantic Web

Eero HYVÖNEN [a,b,1], Hien CAO [a], Rafael LEAL [a], Heikki RANTALA [a], and Aki HIETANEN [c]

[a] *Semantic Computing Research Group (SeCo), Department of Computer Science, Aalto University, Finland*

[b] *University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG), Finland*

[c] *Ministry of Justice, Finland*

ORCiD ID: Eero Hyvönen https://orcid.org/0000-0003-1695-5840, Hien Cao https://orcid.org/0009-0004-2774-8243, Rafael Leal https://orcid.org/0000-0001-7266-2036, Heikki Rantala https://orcid.org/0000-0002-4716-6564

Abstract. This paper addresses the issue of searching legislative documents in an international multilingual setting. Legal documents are published in different countries using local languages, as well as country-specific semantic keyword and classification systems. Consequently, when moving from one country to another, citizens may face obstacles when looking for regulations on immigration, heath care, education, etc. To address this challenge, this paper presents FINESTLAWSAMPO, a cross-border solution approach and proof-of-concept demonstrator based on Linked Open Data (LOD) and Semantic Web technologies. We describe the design and implementation of this idea using consolidated laws from Finland and Estonia alongside EU directives as a case study. The demonstrator includes a LOD service and a semantic portal, based on the Sampo Model, which adapts the interface and contents to the user-chosen language. The main novelty presented is the provision of heterogeneous cross-border, multilingual, distributed legal data through faceted searching and data exploration as well as using data analysis in legal informatics.

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

1. Introduction

Legislation and case law are widely published online in different countries by their governments to make jurisdiction transparent and freely accessible to the public, organizations, and lawyers [1]. In an international setting, such as the EU, cross-border access to legislation published in different countries in different languages is often needed. Even though legislation is often available openly, it is not necessarily Findable, Accessible, Interoperable, and Re-useable according to the FAIR[2] principles for scientific data man-

[1] Corresponding Author: Eero Hyvönen, eero.hyvonen@aalto.fi

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

agement and stewardship.

A specific problem here is that local legislation and user interfaces (UI) for searching and browsing 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[3] [2], e.g., in computational law[4].

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. However, it should be made clear and explicit to the end user when machine translations are used as they they may contain errors. To support this argument, a model for publishing and using cross-border multilingual legislation databases is presented. As a case study, the integration of 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[5] on top of which the portal was created[6]. The data as well as the portal software[7] are available openly online CC BY 4.0.

In the following, related works are first discussed and after that our model for publishing and using cross-border multi-lingual legislation is presented. Then we detail the LOD underlying FINESTLAWSAMPO and how it was created using, for example, natural language processing (NLP). After this, the usage of the data service and the portal on top of it are explained. In conclusion, contributions of the paper and lessons learned are summarized.

2. Related Works

The Web provides a promising medium for publishing legal documents. There are, for example, portals, such as `www.legislation.gov.uk` for the legislation for the UK, Scotland, Wales, and Northern Ireland[8], and EU level systems, such as HUDOC[9], EUR Lex [10], the EU Cellar[11], and the ECLI Search Engine[12] for the case law. There are various legislative info portals which serve both citizens, professionals, and government

[3] `https://en.wikipedia.org/wiki/Legal_informatics`
[4] `https://law.stanford.edu/2021/03/10/what-is-computational-law/`
[5] LOD data service online: `https://ldf.fi/datasets/finestlaw`
[6] Portal online: `https://finestlaw.demo.seco.cs.aalto.fi/en`
[7] `https://github.com/SemanticComputing/finest-lawsampo-web-app`
[8] `https://www.legislation.gov.uk`
[9] `https://hudoc.echr.coe.int/`
[10] `https://eur-lex.europa.eu/`
[11] `https://data.europa.eu/euodp/en/data/dataset/sparql-cellar-of-the-publications-office`
[12] `https://e-justice.europa.eu/content_ecli_search_engine-430-en.do`

agencies, such as EUR-Lex, N-Lex[13], and European e-Justice[14]. Euro-Lex is based on the European Union's (EU) official databases and gives access to EU legislation, case-law, and other legal documents. N-Lex is a legal search interface serving as a gateway to access to national databases of individual EU countries. European e-Justice offers information on the EU justice system and enhances accessibility to justice across the EU, i.e., providing information about finding legal professionals. It mainly provides information about judicial processes, while Euro-Lex and N-Lex offer access to legal texts.

Our work on legal Linked Data services was influenced by the MetaLex Document Server[15] [3] and related national online services for legal documents in Greece, Luxemburg[16], France, Norway[17], and the U.S. [4]. 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[18] in the U.S.

3. A Model for Publishing Cross-border Legislation

The Sampo model [5] and the Sampo-UI framework [6,7] was used 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 [8]; 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 six general principles on how to create 1) LOD services and 2) user interfaces that utilize them. The model has evolved gradually during the development of over twenty LOD services and semantic portals[19], mostly in the domain of Cultural Heritage (CH) and Digital Humanities (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 (1452–1519) has maintained[20]: *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, this applies to the Finnish and Estonian legislation and the related EU directives.
2. *Use a shared open ontology infrastructure (P2)* According to a wisdom[21] of Albert Einstein (1879–1955) *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 [9]: it is arguably better to prevent interoperability problems already when creating data than fix problems afterwards when aggregating data [10]. In our case, it was possible to re-use the LAWSAMPO infrastructure,

[13] N-Lex: `https://n-lex.europa.eu/n-lex/index`
[14] European e-Justice: `https://e-justice.europa.eu/home?action=home&plang=en`
[15] `http://doc.metalex.eu`
[16] `http://legilux.public.lu/editorial/eli`
[17] `http://lovdata.no/eli`
[18] `https://scholar.google.com/scholar_courts`
[19] Sampo systems homepage: `https://seco.cs.aalto.fi/applications/sampo/`
[20] `https://philosophynow.org/issues/134/The_Mind_of_Leonardo_da_Vinci`
[21] `https://elevatesociety.com/intellectuals-solve-problems-geniuses-prevent/`

EU level standards, such as ELI identifiers[22] [11], and controlled vocabularies, especially the multilingual Pan-European EuroVoc thesaurus[23] 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 [12], which aimed at answering the question whether it is possible to implement ontology services [13,14] by using SPARQL queries only for data access. It was also tested whether it makes sense to apply this idea to the implementation of faceted semantic search, as used in Sampo systems since 2004. The answer was positive, and this result led to the development of the tools SPARQL Faceter [15] and later Sampo-UI [6] that has been used in some 15 Sampos including FINESTLAWSAMPO.

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

1. *Provide multiple perspectives to the same data (P4)* The idea here is the same as in the FAIR principles[24], but adapted to UI design: reusing the data even within one UI. The class structure of Knowledge Graphs (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 prosopographical research method on groups of people [16], where a target group 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 [17].

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 the 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. The burden of actually executing the query can also be distributed. Moreover, central-

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

[23] https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/eurovoc

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

ized federated query processing can also be used. This is possible, e.g., in SPARQL, but 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[25] 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 different local datasets is hard.

In the Sampo model used in FINESTLAWSAMPO the centralized strategy was therefore selected, introduced already in the first Sampo system MuseumFinland [18] (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,the ways they are created, and have to be addressed in any case when dealing with local data in a semantically proper interoperable way.

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 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 [19], vocabularies used in populating the models, and the data itself evolves. This problem is discussed, e.g., in [20,21].

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 here is that the new data is likely to contain typos and linking textual descriptions may need manual work and fixes after all. For detecting quality issues, semantic validation languages and frameworks, such as SHACL[26] and ShEx[27] 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[28] and Topbraid Composer[29]. In the case of

[25] 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.

[26] `https://www.w3.org/TR/shacl/`

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

[28] `https://protege.stanford.edu/`

[29] `https://allegrograph.com/topbraid-composer/`

the Sampo systems, the SPARQL SAHA editor [22] was developed and has been used in maintaining Sampo datasets by their data owners in some cases.

4. Creating the Knowlegde Graph and LOD Service

This section overviews the data underlying FINESTLAWSAMPO, how it was transformed into a KG, and published as a LOD service.

4.1. Primary Data and Data Model

Finnish legislation and case law decisions have been published as web documents since 1997 in the Finlex Data Bank[30]. 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 [23] 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. This data model was re-used in FINESTLAWSAMPO.

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[31]. 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[32] in Estonian and in English. The XML documents were transformed into the LAWSAMPO RDF data model using Python SAX APIs [33] and Python RDFLib [34].

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

[30] `http://www.finlex.fi`
[31] `https://www.dublincore.org/specifications/dublin-core/dcmi-terms/`
[32] Consolidated texts of of Estonian legislation: `https://www.riigiteataja.ee/en/`
[33] `https://docs.python.org/3/library/xml.sax.handler.html`
[34] `https://github.com/RDFLib/rdflib`

4.2. Data Enrichment

The original datasets were enriched by Data Linking. In order to further enrich the legal document data, several NLP techniques were used: the contents were translated automatically to cover all wanted languages (Finnish, Estonian, English); keywords were extracted; the documents were classified using two different sets of Life Event categories used in public services in Finland and Estonia, and their similarity was calculated.

4.2.1. Internal and External Linking

The data was linked internally to improve the references to related documents and their parts. A challenge here is that specific earlier versions of a statute may be referred to.

Both Finnish and Estonian statutes were linked externally to EU directives that were mentioned in them. The directives were available at the EU Cellar[35]. 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 format.

4.2.2. 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, i.e., 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 were missing. We decided to use automatic machine translation to fill this gap, even though these translations do not have legal force.

The translations were carried out using the Opus-MT's machine translation models [24] of the University of Helsinki. These are a series of deep learning translation models, each one fine-tuned for a specific language pair. We used the corresponding models[36] via the Huggingface's Transformers library [25].

These models typically have a maximum input length of 512 tokens, which is roughly equivalent to 400 words in English or 300 in Finnish. This means that only short text snippets can be translated at once, which means that the whole of the context cannot 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.

In order to preserve the original HTML format in the translations, the Python library XML2Dict[37] was used to unpack and repack the sentences.

4.2.3. Keyword extraction

The objective of extracting keywords for each document is twofold: on the one hand, it allows the users of the portal to navigate documents succinctly and conveniently by

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

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

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

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[38] 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* [26], which behaves as a document classification tool over a pool of limited and known possible keywords.

4.2.4. Classification Based on Life Situations

A novelty of FINESTLAWSAMPO is to provide the end user the statues from a perspective of life situations in which the end user is expected to be involved. Both in Finland[39] and in Estonia[40], 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

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

In the FINESTLAWSAMPO KG, all legal documents were automatically classified into both Finnish and Estonian life event 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 the same set.

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

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

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

In order to carry out the classification, the EuroVoc keywords assigned to each document are used as the basis for their internal representation. The keywords are translated to Finnish where needed and, since they are used without any context, transformed into embeddings using fastText [27] instead of state-of-the-art context-aware models based on the Transformers architecture. The embeddings are then pooled together to form the document representation. Translating to Finnish is important in order to allow for semantic reinforcement of the categories using the General Finnish Ontogoly YSO[41] [28] as well as for measuring text similarity, for which simple cosine similarity calculation was used. More details about this process can be found in [28].

This classification system is unsupervised and zero-shot, i.e., it does not use similar examples for training, which makes it nimble. However, the technologies used are not highly performant, which reinforce the proof-of-concept nature of FINESTLAWSAMPO. As an informal makeshift evaluation, we collected Estonian laws mentioned in the "Citizens" section of the Estonian legislation portal Riigiportaal[42], bypassing intermediate life event levels. Each of these laws, which amount to 96, may be associated with multiple life situations. When tested against this golden standard, our classification system scored 0.348 when taking into account only the highest-ranked category assigned by our system (Hits@1), and 0.489 when considering all the proposed categories.

4.2.5. Linked Open Data Service

The FINESTLAWSAMPO data service adopts the 5-star Linked Data model[43], extended with two more stars, as suggested in the Linked Data Finland model and platform [29]. The 6th star is obtained by providing the dataset schemas and documenting them. The FINESTLAWSAMPO schema can be downloaded from the service[44] and the data model is documented using the LODE service[45]. 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 data model used and its integrity constraints are presented in a machine-processable format using the ShEx Shape Expressions language[46] [30]. We have made initial validation experiments with the PyShEx[47] 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 LOD service is powered by the Linked Data Finland[48] 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[49]. As the triplestore, Apache Jena Fuseki[50] is used as a Docker container, which allows ef-

[41] `https://finto.fi/yso/en/`
[42] `ttps://www.eesti.ee/en`
[43] `https://www.w3.org/DesignIssues/LinkedData.html`
[44] `https://www.ldf.fi/dataset/lawsampo`
[45] `https://essepuntato.it/lode/`
[46] `https://shex.io`
[47] `https://github.com/hsolbrig/PyShEx`
[48] `http://ldf.fi`
[49] `https://ldf.fi/lawsampo/sparql`
[50] `https://jena.apache.org/documentation/fuseki2/`

ficient provisioning of resources (CPU, memory), portability, and scaling. Varnish Cache web application accelerator[51] is used for routing URIs, content negotiation, and caching.

5. 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[52]. 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[53]. 4) Most importantly, the data service can be queried in flexible ways using the SPARQL query language[54] and endpoint. There are easy to use tools, such as Yasgui [31], 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 [32,33] 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" [34,35] and also as "dynamic taxonomies" [36].

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 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 [37,38]. It is possible to find potentially interesting subsets and individuals in the

[51] `https://varnish-cache.org`
[52] See, e.g., the browser for DBpedia: `https://dbpedia.org/ontology/Browser`
[53] Content Negotiation by Profile: `https://www.w3.org/TR/dx-prof-conneg/`
[54] SPARQL 1.1 Query Language: `https://www.w3.org/TR/sparql11-query/`

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.

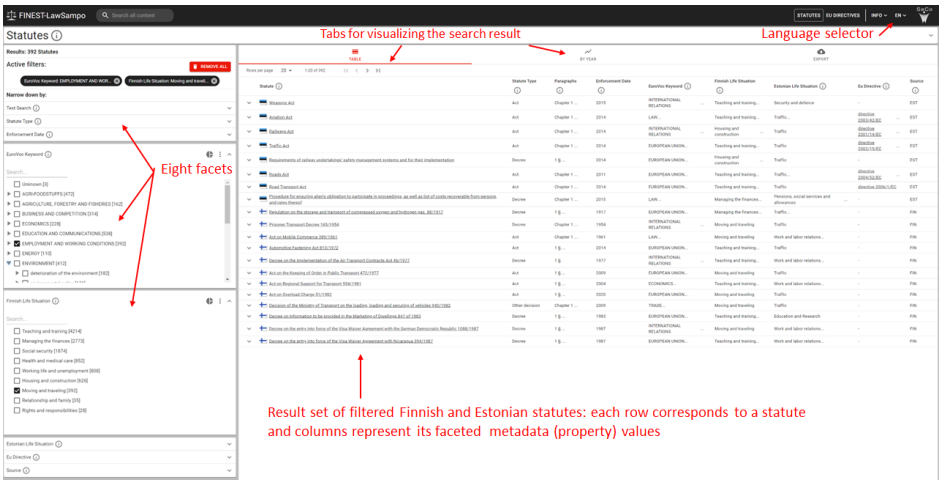

Figure 1. Faceted search for Finnish and Estonian statutes

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 Travelling. The statutes found are show on TABLE tab on the right; country flags show their origin.

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 can 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).

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) the 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 user 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

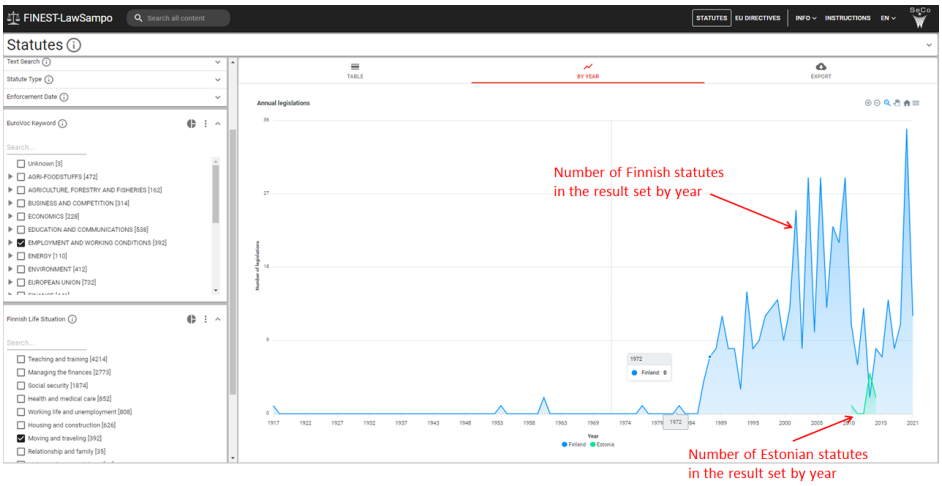

Figure 2. Visualizing the number of statutes in the result set on a timeline by their enforcement day

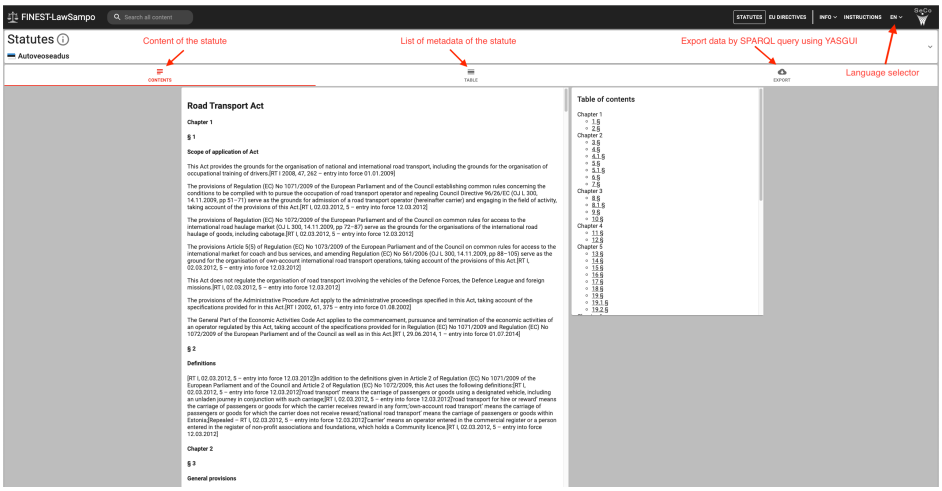

Figure 3. Homepage of a single statute with tabs

SPARQL editor is opened with a query for retrieving the statute(s) of the result set. This feature, available also at the main application perspective page (Fig. 1), is provided for users interested in learning how to query the data by themselves.

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 an 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),

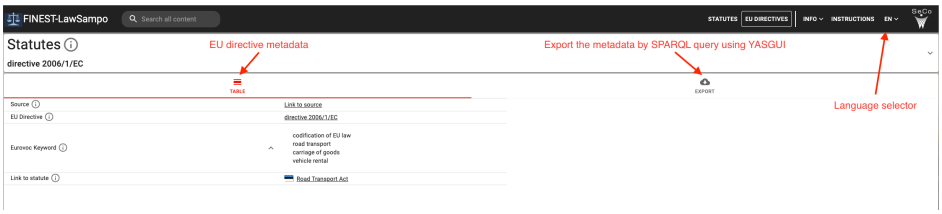

Figure 4. Homepage of a single EU directive, including a link to full source data in EU Cellar

2) EuroVoc keywords, and 3) Links to the Finnish and Estonian statutes mention the EU directive in their textual contents.

6. Comparison with N-Lex

Given FINESTLAWSAMPO's focus on searching national legislation in different countries and languages, it makes sense to compare it with N-Lex that has the same focus.

N-Lex was developed with federated strategy using local national legal web services that utilize different standards and methods for data indexing and classifications. This approach presents challenges for semantically querying or analyzing the data and affects the precision and recall of information retrieval. In contrast, FINESTLAWSAMPO uses a centralized approach, harmonizing and aggregating local data, including human-made and machine translations, into a global knowledge graph. This allows for the creation of semantically interoperable, accessible, high quality content based on ontologies. This approach is arguably able to return much better search results and enable data-analyses.

Regarding features and functionalities, FinEstLawSampo provides novel features such as faceted semantic search on texts, type of legislation, year of publication, EuroVoc keywords, and country based life events. The data has been enriched using data linking, keyword extraction, topical classification, and machine translation, and search is intergrated with data visualization and SPARQL querying. The portal focuses on providing a user-friendly interface offering multiple languages and offering legal data as texts and metadata to support diverse groups of users and application developers. In comparison, N-Lex has a user interface that supports multi-lingual text search with translation of results into national legislation languages. The translation of queries is based on EuroVoc thesaurus and machine translation is used for the result documens. N-Lex also offers the multi-country search which is convenient if someone want to search several national legislative data simultaneously. However, N-Lex is limited in its search capabilities, primarily focusing on text search that is ambiguous in many ways, type of legislation, and year of publication. Furthermore, since N-Lex uses different countrys' databases, the quality of search results varies and the connection to a database may be down.

Serving as a gateway to national legislation of each EU country, N-Lex aims to provide legislative information tailored for legal experts. It supports those who need to read the original legislative texts, compare legislation between EU countries, and find out connections between EU and national legislations. In contrast, FINESTLAWSAMPO provides the service for wider audiences, i.e ., citizens, businesses, researchers, and legal professionals. It not only offers search based on EuroVoc keywords and national life event classifications, but also offers rich metadata for applications,including links to

the original legislative texts, type of legislation, year of publication, EuroVoc keywords, life event categories, related EU directives, and similar legislation in different countries.

FINESTLAWSAMPO serves as a proof-of-concept system that utilizes the legislation data of Finland, Estonia, and EU to demonstrate cross-border multi-lingual legal data access. In contrast, N-Lex is an in-use legal portal that provides broader scope of coverage by offering access links to legislative databases across 27 EU countries. However, FinEstLawSampo enriches the data and stores it in a single triplestore with APIs and an open SPARQL endpoint for FAIR data, while N-Lex doesn't contain documents but only provides federated search and links to national legislative databases.

Regarding implementation, maintenance, and sustainability, N-Lex is simpler system as maintaining the data and querying services are distributed to local databases and data providers, while in FINESTLAWSAMPO the data is harmonized and aggregated into a global triplestore and used by a centralized UI.

7. Discussion

This paper applied the Sampo Model, developed originally for Digital Humanities research, to a novel use case in multi-lingual cross-border legal informatics. Here legal documents from different countries written in different languages are harmonized using shared ontologies and a data model. The texts are automatically translated and enriched with contextual linked data using data linking and NLP techniques. As a use case, the FINESTLAWSAMPO demonstrator was presented where the end user is provided with ready-to-use faceted multi-lingual search and browsing with data-analytic tools for analyzing the documents.

It was argued, that centralized approach studied in our work is able to provide a way to create semantically interoperable rich data for intelligent applications and data-analyses. However, it is is more complicated to implement and maintain than the traditional distributed approach based on federated search.

FINESTLAWSAMPO has access only to the latest versions of manually consolidated statues available in Finlex, and retrieves Estonian legislation database in Riigiteataja[55] 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 [39] 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 [5].

An informal evaluation of the portal suggests that relevant legislation can be found without lots of error, but a more accurate evaluation is challenging. There is no gold standard available and determining precision and recall in a search task would require substantial international legal expertise. It should be noted that the portal also includes a traditional text search facet an option to use, and from that perspective it is more versatile than 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 users are probably interested in exploring legislation data they are not familiar with.

[55] `https://www.riigiteataja.ee/avaandmed/ERT/`

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, agreements and standards between the national legal data publishers.

Regarding NLP techniques and models used in this system, it is important to emphasize its proof-of-concept nature. The technologies used are not the most advanced, since the emphasis lied in a fast and lightweight implementation of the portal. The integration of state-of-the-art Large Language Models (LLMs) could be beneficial, although not without caveats. The first, and arguably most important, is multi-lingual support: few of the largest LLMs are able to operate in Finnish with ease, not to mention Estonian. The integration of yet other languages could prove to be a challenge. The arguably best LLMs in this regard, such as GPT-4 or Gemini, tend to be closed-source, which is another obfuscation step in addition to the black-box Transformer architecture. If control of the system is desirable, the most capable LLMs are inadequate. Hallucinations may also be a concern both in tranlation and classification.

On the other hand, when it comes to text classifications, LLM embeddings are certainly much more performant than fastText's, since they are trained on updated algorithms and with much larger datasets. Moreover, computational overhead of LLMs are probably not prohibitive for this use case, since the computations can be performed beforehand. As a result, it would be possible to delegate both translation and classification to a single LLM or to extract its embeddings for classification elsewhere.

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] van Opijnen M, Peruginelli G, Kefali E, Palmirani M. Online Publication of Court Decisions in Europe. Legal Information Management. 2017;17:136—145.

[2] Erdelez S, O'Hare S. Legal Informatics: Application of Information Technology in Law. Annual Review of Information Science and Technology. 1997 01;32.

[3] Hoekstra R. The MetaLex Document Server Legal Documents as Versioned Linked Data. In: Proceedings of the ISWC 2011. Springer; 2011. p. 128-43.

[4] Casellas N, Bruce TR, Frug SS, Bouwman S, Dias D, Lin J, et al. 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. p. 280-1.

[5] Hyvönen E. Digital Humanities on the Semantic Web: Sampo Model and Portal Series. Semantic Web – Interoperability, Usability, Applicability. 2022:1-16. Available from: `https://doi.org/10.3233/SW-223034`.

[6] 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. 2022 January;13(1):69-84.

[7] 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. Available from: `https://ceur-ws.org/Vol-3508/paper3.pdf`.

[8] Hyvönen E, Tamper M, Oksanen A, Ikkala E, Sarsa S, Tuominen J, et al. 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. p. 110-4.

[9] 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. Under review. Available from: `https://seco.cs.aalto.fi/publications/2022/hyvonen-infra-2022.pdf`.

[10] Hyvönen E. Preventing interoperability problems instead of solving them. Semantic Web – Interoperability, Usability, Applicability. 2010;1(1–2):33-7.

[11] 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. Publications Office of the EU; 2012. p. 3-11.

[12] 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. Available from: `https://seco.cs.aalto.fi/publications/2012/suominen-et-al-onkilight-2012.pdf`.

[13] 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. .

[14] 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. Springer; 2009. p. 781-95.

[15] 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, editors. Joint Proc. of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop. CEUR Workshop Proceedings, Vol-1615; 2016. Available from: `http://ceur-ws.org/Vol-1615/semdevPaper5.pdf`.

[16] Verboven K, Carlier M, Dumolyn J. A short manual to the art of prosopography. In: Prosopography approaches and applications. A handbook. Unit for Prosopographical Research (Linacre College); 2007. p. 35-70.

[17] 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. 2020;11(1):187-93.

[18] Hyvönen E, Mäkelä E, Salminen M, Valo A, Viljanen K, Saarela S, et al. MuseumFinland—Finnish Museums on the Semantic Web. Journal of Web Semantics. 2005;3(2):224-41.

[19] Zeng M, Qin J. Metadata, Third Edition. ALA Neal-Schuman, Chicago; 2022.

[20] 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.. vol. 11196 of Lecture Notes in Computer Science. Springer; 2018. .

[21] 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. Available from: `http://ceur-ws.org/Vol-2180/`.

[22] 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. .

[23] 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. IOS Press; 2019. p. 212-28.

[24] 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, et al., editors. Proceedings of the 22nd Annual Conference of the European Association for Machine Translation. European Association for Machine Translation; 2020. p. 479-80. Available from: `https://aclanthology.org/2020.eamt-1.61`.

[25] 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. Association for Computational Linguistics; 2020. p. 38-45.

[26] Avram AM, Pais V, Tufis DI. 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). INCOMA Ltd.;. p. 92-101. Available from: https://aclanthology.org/2021.ranlp-1.12.

[27] Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics. 2017 12;5:135-46.

[28] 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). vol. 93 of Open Access Series in Informatics (OASIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik; 2021. p. 18:1-18:15.

[29] 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. Springer; 2014. p. 226-30.

[30] Thornton K, Solbrig H, Stupp GS, Gayo JEL, Mietchen D, Prud'hommeaux E, et al. Using Shape Expressions (ShEx) to Share RDF Data Models and to Guide Curation with Rigorous Validation. In: The Semantic Web. ESWC 2019. Springer; 2019. p. 606-20.

[31] Rietveld L, Hoekstra R. The YASGUI family of SPARQL clients. Semantic Web – Interoperability, Usability, Applicability. 2017;8(3):373-83.

[32] Hearst M. Design recommendations for hierarchical faceted search interfaces. In: ACM SIGIR workshop on faceted search. Seattle, WA; 2006. p. 1-5.

[33] Tunkelang D. Faceted Search. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool, Palo Alto, CA, USA; 2009.

[34] Pollitt AS. The key role of classification and indexing in view-based searching. UK: University of Huddersfield; 1998. Available from: http://www.ifla.org/IV/ifla63/63polst.pdf.

[35] 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. .

[36] Sacco GM. Dynamic taxonomies: guided interactive diagnostic assistance. In: Wickramasinghe N, editor. Encyclopedia of Healthcare Information Systems. Idea Group; 2005. .

[37] Marchionini G. Exploratory search: from finding to understanding. Communications of the ACM. 2006;49(4):41-6.

[38] Tzitzikas Y, Manolis N, Papadakos P. Faceted exploration of RDF/S datasets: a survey. Journal of Intelligent Information Systems. 2017;48(2):329-64.

[39] Burrows T, Pinto NB, Cazals M, Gaudin A, Wijsman H. Evaluating a Semantic Portal for the "Mapping Manuscript Migrations" Project. DigItalia. 2020;2:178-85. Available from: http://digitalia.sbn.it/article/view/2643.