Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190023

Semantic Finlex: Transforming, Publishing, and Using Finnish Legislation and Case Law As Linked Open Data on the Web

Arttu OKSANEN [a,b], Minna TAMPER [b], Jouni TUOMINEN [b]
Eetu MÄKELÄ [b], Aki HIETANEN [c] and Eero HYVÖNEN [b]
[a] *Edita Publishing Ltd. (Finland)*
[b] *Semantic Computing Research Group (SeCo), Aalto University (Finland)*
[c] *Ministry of Justice of Finland*

Abstract. Governments publish legislation and case law widely in print and on the Web. Such legal information is provided for human consumption, but the information is usually not available as data for algorithmic analysis and applications to use. However, this would be beneficial in many use cases, such as building more intelligent juridical online services and conducting research into legislation and legal practice. To address these needs, this Chapter presents Semantic Finlex, a national in-use data resource and service for publishing Finnish legislation and related case law as Linked Open Data for legal applications to use. The system transforms and interlinks on a regular basis data from the legacy legal database Finlex of the Ministry of Justice into Linked Open Data, based on the European standards ECLI and ELI. The published data is hosted on the '7-star' Linked Data Finland service and SPARQL endpoint with a variety of related services available that ease data re-use. Rich Internet Applications using SPARQL for data access are presented as application demonstrators of the data service. In addition, this Chapter presents methods and tools under development to automatically annotate legal texts and to anonymize case law documents prior to their publication on the Web. Anonymization is necessary due to issues of data protection and privacy, and annotation is needed for semantic search and interlinking the documents. The automated approaches could significantly speed up the process and minimize costs of publishing legal documents as Linked Open Data.

Keywords. legislation, case law, linked data publishing, automatic anonymization, automatic annotation

1. Introduction

Governments provide publicly available legal information on the Web usually in the form of HTML or PDF documents targeted to human readers. In Finland, for example, legislation and case law are published as HTML documents in the Finlex Data Bank[1], a

[1] http://www.finlex.fi.

publicly available online service since 1997, maintained by the Ministry of Justice [1]. However, Finlex does not provide publicly available machine-readable legal information as open data, on top of which services and analyses can be built by the ministry or third party vendors.

This Chapter presents Semantic Finlex[2], a national Linked Open Data Service for Finnish legislation and case law. The service hosts and publishes a central part of the Finnish legislation along with judgments of the Supreme Court and the Supreme Administrative Court. All of the datasets are automatically updated regularly.

Our work on Semantic Finlex started in 2012, and the first version of the service was published in 2014 [2]. The data included 2,413 consolidated laws, 11,904 judgments of the Supreme Court, and 1,490 judgments of the Supreme Administrative Court. In addition, some 30,000 terms used in 26 different thesauri were harvested for a first draft of a consolidated vocabulary. During this work, some shortcomings of the initial RDF data model became evident as well as the need for using the then emerging new standards for EU level interoperability. The demo dataset also consisted of only one temporal version (2012) of the statutory law and was not updated. These issues have now been resolved in the work reported in this article.

In the following, we first explicate the motivation and use cases for publishing legislation and case law as linked open data. Then the underlying data models and the data conversion process applied in the service are presented, followed by a discussion on enriching the data with semantic and structural annotations. Then, in Section 5, we introduce the Semantic Finlex publishing platform and semantic portal. In Section 6, data analysis and application demonstrators built on top of the service are presented. Finally, in Section 7 we present our ongoing work to automatically anonymize and annotate legal documents.

2. Motivating Use Cases

Many actors and tasks would benefit from access to legislative and judicial content as data:

Information portals. Within the online services provided by different sectors, it is often necessary to refer to various sections of acts and decrees and display these to users. This requires that such sections be referable and readable as online data. For example, various regulations referring to law are published in the fields of construction, defense, and chemical safety.

The media. Since news on fields such as politics and the business world often refer to various sections of statutes, it is sometimes useful to guide readers to the original legal texts. However, this is not possible if the sections in question are not referable or available in data format.

Juridical online services. In Finland, these include services such as Suomen Laki (Finnish Law)[3] by Talentum Oyj and Edilex[4] by Edita Publishing Ltd, which primarily provide juridical information for professionals in law, such as judges and legal counsels, as well as private persons. Maintaining data in current systems is tedious and largely

[2] http://data.finlex.fi.
[3] http://www.suomenlaki.com.
[4] http://www.edilex.fi.

based on manual work, because the data is not available in a form 'understood' by computers, but only as documents in PDF, Word, and other formats.

Legislative drafting. When new statutes are drafted in order to complement and supersede previous ones, the drafters have to examine previous statutes in order to evaluate the effects of the changes and avoid discrepancies. However, semantic information on the various versions of and interdependencies between statutes has been available only in text format.

Editing and publishing of legislative texts. Today, legislation-related information is produced in an inconsistent manner, by using various text formats and index term vocabularies to describe information content. If documents were drafted at the production stage in the form of structured data and in accordance with mutually agreed standards, this would facilitate their further processing and linking to other documents, such as materials in Parliament and in publishing systems such as Finlex.

Intelligent services. Legislative information related to problematic juridical situations, such as divorce or estate distribution, is often scattered between various acts, decrees, and legal practice cases. The availability of statutes and legal cases as such is of little help if the reader, such as an ordinary citizen, finds it impossible to piece the issue together. Presenting legislative documents in a form that can be interpreted by a computer, i.e., as semantic data, would enable the development of more intelligent applications, which would in turn enable making law and justice more comprehensible to citizens. For example, legal texts can be automatically linked to other related texts, legal cases, and vocabularies explaining legal terminology.

Research into legislation and legal practice. The enactment of legislation and legal practice are fields of research in which data analysis methods can be used. The topic of such a research might, for instance, be the impact of EU law to national legal practice [3]. However, data analysis methods require that statutes, the connections between them, and case law-based information on their implementation are available in the form of systematically presented data.

Moreover, authorities in Europe strive to improve the semantic interoperability between EU and Member State legal systems, as the methods in use now for storing and displaying legal documents differ among countries. Therefore, the Council of the European Union has invited the introduction of ELI (European Legislation Identifier)[5] and ECLI (European Case Law Identifier)[6] standards that define common identifier and metadata models for legislative and case law documents by applying Linked Data principles.

3. Conversion to Linked Data

This Section presents the datasets, data models, and the conversion process in use in the Semantic Finlex data service.

[5]Council of the European Union (2012). Council Conclusions Inviting the Introduction of the European Legislation Identifier (ELI). *Official Journal of the European Union*, C 325, 3-11.

[6]Council of the European Union (2011). Council Conclusions Inviting the Introduction of the European Case Law Identifier (ECLI) and a Minimum Set of Uniform Metadata for Case Law. *Official Journal of the European Union*, C 127, 1-7.

3.1. Datasets

The Semantic Finlex service currently consists of four different datasets: *Original legislation.* This dataset consists of approximately 49,000 acts and decrees as they originally appeared in the Statutes of Finland, the official publication of Finnish Law. Besides new acts and decrees, the dataset includes amendments and repeals targeted on previously enacted statutes.

Consolidated legislation. Consolidated texts of acts and decrees incorporate their successive amendments. Editorial work has been carried out by a publishing company. Currently, the dataset includes approximately 3,100 statutes.

Judgments of the Supreme Court. This dataset comprises approximately 5,500 precedents published in the Yearbook of the Supreme Court since 1980.

Judgments of the Supreme Administrative Court. This dataset includes roughly 7,500 judgments of the Supreme Administrative Court from 1987 onwards.

All of the datasets are transformed from different legacy XML formats to RDF adapting the ELI and ECLI specifications. New and updated documents are fetched weekly from the Finlex service and converted into RDF.

3.2. Data Models

As law is not constant but changes over time the RDF data model needs to be able to identify the different temporal versions of the law. Secondly, statutory citations refer to both entire statutes and their individual parts. Therefore, we need to identify the different document parts as well as their temporal versions. Moreover, different language versions of the same document may exist and content can be represented in multiple formats that all need to be identified.

The ELI standard applies the well-established conceptual model FRBR (Functional Requirements for Bibliographic Records)[7] to its ontology definition to distinguish between (1) statutes as such (*work*), (2) their different versions (*expression*), and (3) different content formats (*manifestation*). As the ELI implementation guide[8] states that in multilingual environments, such as in Finland where both Finnish and Swedish are official languages, expressions should be used only to model the different language versions, we use the work level to model temporal versions.

We extended the ELI ontology with our own ontology named SFL (Semantic Finlex Legislation) to define separate classes for the different work level entities, i.e., the legislative documents and the document parts as such (*sfl:Statute* and *sfl:SectionOfALaw*), and their temporal versions (*sfl:StatuteVersion* and *sfl:SectionOfALawVersion*). The extended data model is presented on the left side of Figure 1. Different namespaces and prefixes used in the schemas are listed in Table 1. Statutes and their parts are linked to their temporal versions using the property *eli:has_member*. Temporal versions in turn have two language variants in Finnish and Swedish. These language variants are modeled as instances of the *eli:LegalExpression* class and linked to the temporal versions using the property *eli:realizes*. Finally, the different content formats (text and HTML) of

[7] IFLA Study Group on the Functional Requirements for Bibliographic Records (1998). *Functional requirements for bibliographic records*. K.G. Saur Verlag.

[8] ELI Task Force (2015). *ELI: A Technical Implementation Guide*. Publications Office of the European Union.

the language variants are modeled as instances of the *eli:Format* class and linked to the language variant using the property *eli:embodies*.

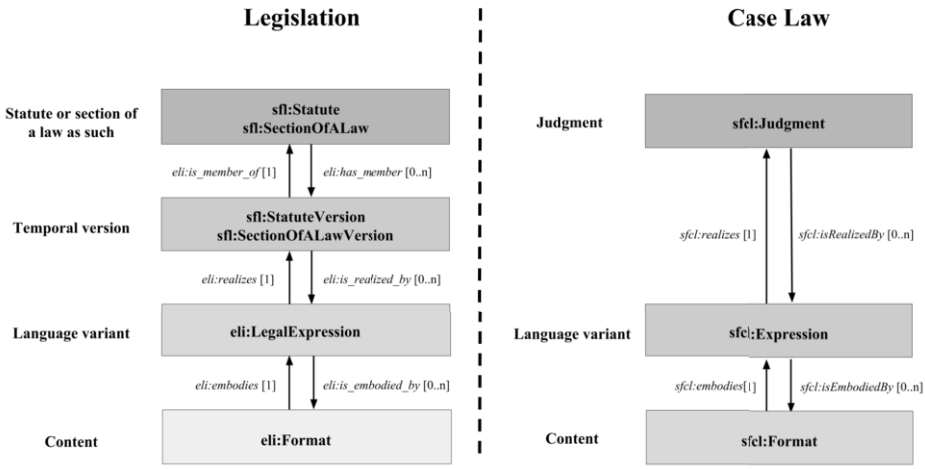

Figure 1. The FRBR-inspired data models for legislative and case law documents

Table 1. Prefixes and namespaces used in the RDF data models

Prefix	Namespace
common	http://data.finlex.fi/common/
dcterms	http://purl.org/dc/terms/
eli	http://data.europa.eu/eli/ontology#
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
sfcl	http://data.finlex.fi/schema/sfcl/
sfl	http://data.finlex.fi/schema/sfl/
skos	http://www.w3.org/2004/02/skos/core#

SFL further extends the ELI ontology with an additional property *sfl:statuteType* to describe the functionality of a statute, i.e., whether it is a new statute, an amendment, or a repeal. ELI itself defines descriptive properties *eli:type_document*, that we use to describe the level of a statute in the hierarchy of norms (i.e., whether the statute is an act, a decree, or a decision), and *eli:version* to distinguish between original and consolidated document versions.

The physical structure of a statute is modeled at the temporal version level, as it can vary between different temporal versions of the same statute. The properties *eli:has_part* and *eli:is_part_of* are used to model the document tree. Each type of section of a law is modeled as a class in the SFL ontology. The model of the physical structure is presented in Figure 2. As an example, a statute might consist of multiple sections, while sections consist of subsections and subsections consist of paragraphs.

For case law, we adapted the ECLI standard. In contrast to the sophisticated functional model of ELI, ECLI only defines a set of Dublin Core properties to be used to

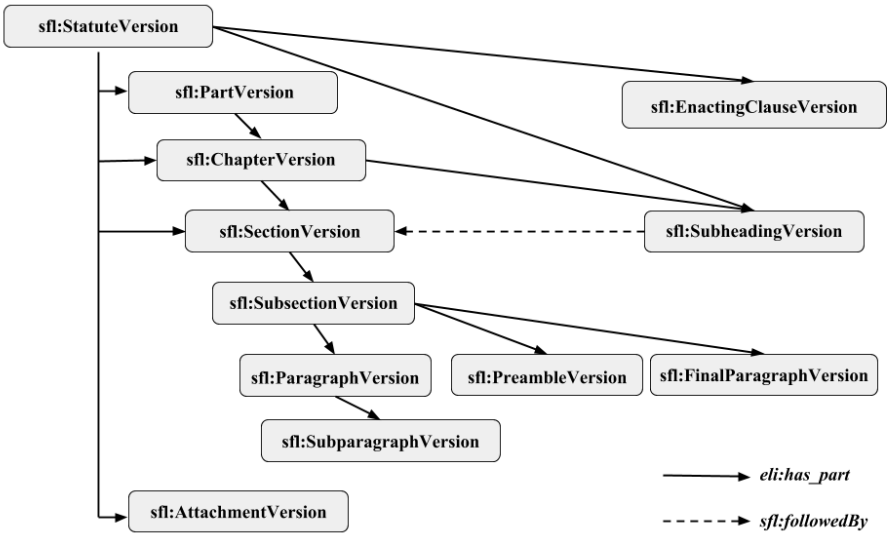

Figure 2. Model of the physical structure of a statute

annotate case law documents. Thus we have developed our own ontology called SFCL (Semantic Finlex Case Law) that wraps the ECLI metadata model to an FRBR-inspired model reminiscent of ELI. However, we can omit temporal versions from the model since there is only one temporal version of each judgment in both Finnish and Swedish. The FRBR model for case law is presented on the right side of Figure 1. Converting the judgments into RDF is quite straightforward as most of the metadata fields mentioned in the definition of ECLI are included in the source data XML.

Contents of both case law and legislative documents are stored at the manifestation level in text and HTML formats as values to RDF properties *sfl:text*, *sfl:html*, *sfcl:text* and *sfcl:html*. This allows the HTML or text content of a specific judgment or section of a law to be retrieved from the triple store with a single SPARQL query.

3.3. URI Identifiers

Besides an ontology, ELI defines a URI pattern schema to unambiguously identify legislation. The URI patterns developed for the Semantic Finlex are presented in Table 2. The original versions are denoted with parameter *alkup* in the URI and consolidated versions with *ajantasa/{date}*, where the date corresponds to the date of entry into force. The documents as well as their parts, their temporal versions, language versions, and content formats can all be identified uniquely using these patterns. As an example, Finnish language version of the Criminal Code of Finland (39/1889) chapter 2 c section 4 as it was in force on February 1, 2018 can be accessed using the URI

http://data.finlex.fi/eli/sd/1889/39/ajantasa/20180201/luku/2c/pykala/4/fin

As for the case law documents, we generate URIs that mimic the ELI pattern, because the standard format for document identifiers defined by ECLI is not an HTTP URI. The URI pattern is presented in Table 3. For example the ECLI identifier

ECLI:FI:KKO:2016:1

is transformed to

http://data.finlex.fi/ecli/kko/2016/1

The document tree structure of the case law documents is not modeled in RDF, and therefore no identifiers for the document parts need to be generated.

Table 2. ELI-compliant URI patterns for legislative documents

URI pattern	Description
/eli/sd/{year}/{id}	Statute as such
/eli/sd/{year}/{id}/pykala/{section id}/...	Section of a law as such
/eli/sd/{year}/{id}/.../alkup	Original (official) version
/eli/sd/{year}/{id}/.../ajantasa/{date}	Consolidated version
/eli/sd/{year}/{id}/.../{version}/{language}	Language variant
/eli/sd/{year}/{id}/.../{version}/{language}/{format}	Content

Table 3. ELI-mimicing URI patterns for case law documents

URI pattern	Description
/ecli/{court}/{year}/{id}	Judgment
/ecli/{court}/{year}/{id}/{language}	Language variant
/ecli/{court}/{year}/{id}/{language}/{format}	Content

3.4. Collecting Version History

To enable access to previous versions of in-force legislation, a version history of consolidated acts and decrees is required. To collect such a history that is queryable by date we need the dates of entry into force of individual statutes and their parts. Resolving the date of entry into force is not always straightforward, since the legacy XML documents do not provide these dates as explicit metadata. Therefore the information must be extracted from the document text using regular expressions which is prone to error.

3.5. Data Validation

Characteristic to the legislative XML formats is the lack of ELI-compliant metadata. In order to produce the RDF metadata it must be extracted from the document text during the conversion process using regular expressions. If the required text is not in the assumed place in the assumed form, then the conversion process may result in missing values and possibly errors in the converted data. Therefore, rules need to be defined to validate the data before allowing it to be published in the service. These rules can be expressed in the form of SPARQL queries performed against the converted RDF data. An example of such a query is one that verifies that the date of entry into force does not precede the date of publication of a statute.

In addition, the ELI validator[9] developed by Sparna Labs can be used to check the conformance of the RDF data against the ELI ontology. The validator applies predefined SHACL shapes[10] (RDF Shapes Constraints Language) to the input data and files a report on the identified constraint violations.

4. Enriching the Data with Relationships and Key Concepts

This Section discusses enriching the data with relationships between different datasets and key concepts describing the contents of the documents.

4.1. Relationships

To describe the interconnectedness of the law, references to different sources of law are extracted from the documents. These relationships are always linked to the most detailed part of the document text.

First of all, the relationship between the consolidated versions and the original amendments is modeled by linking the amendments to the corresponding consolidated versions with the *eli:amends* and *eli:amended_by* properties. This model is depicted in Figure 3. If the section of a law in question has been repealed, a link to the repealing statute is created instead, as presented in Figure 4.

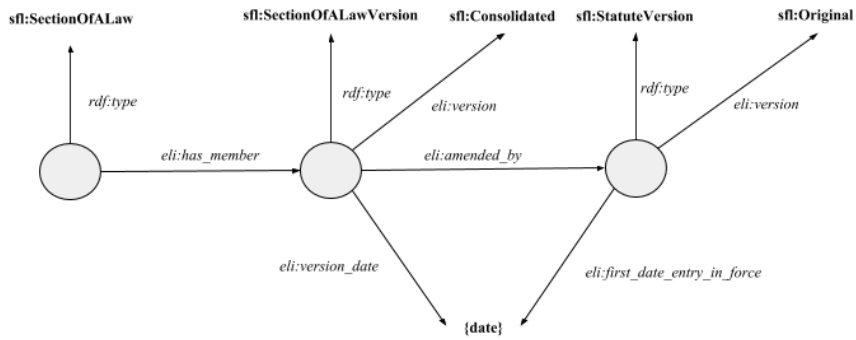

Figure 3. Consolidated versions are linked to the corresponding original amendments.

In addition, references to both national and EU legislation are extracted from the statutes. References to national legislation are denoted using the property *eli:cites*. ELI defines the *eli:transposes* property to describe which legal acts of the EU a statute transposes into national law. The original versions of the statutes contain references to EU directives and regulations for which ELI-compliant URIs can be generated automatically by following the ELI URI pattern.

References to legislative texts are also extracted from case law documents. These are annotated with the property *sfcl:referenceToLegislation*. However, since we do not

[9] http://labs.sparna.fr/eli-validator/.
[10] http://www.w3.org/TR/shacl/.

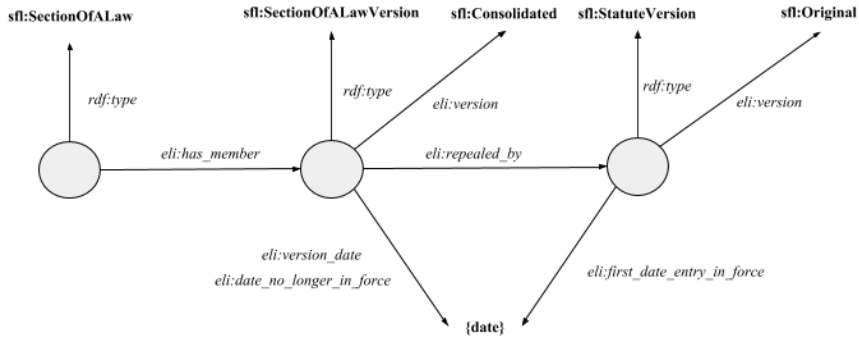

Figure 4. Repealed consolidated versions are linked to the corresponding original repeals.

know which version of a legislative text the citation refers to, we always resolve the link to the abstract work level of a statute or a section of a law, and not any specific temporal version.

To further enrich the metadata of the case law documents, names of the justices of the Supreme Court are extracted from the texts. This is done by using regular expressions that match known types of names. The personnel are modeled as *dcterms:Agent* type of resources and linked to a specific judgment with the property *dcterms:contributor* in accordance with the ECLI specification.

4.2. Key Concepts

To support search and discovery of legal texts, key concepts relating to pieces of legislation were automatically mined from the texts of the documents [4]; [5]. These semantic annotations were selected from the following vocabularies: The Bank of Finnish Terminology in Jurisprudence[11], Eurovoc[12], the legal terminology sections of the KOKO ontology[13], and DBpedia[14].

Before querying the vocabularies, the texts were filtered using stopword lists and linguistic tools. First, the entire text was lemmatized using the SeCo Lexical Analysis Services [6]. The lemmatized results were first filtered based on part-of-speech tags (accepting only words and compound words with proper nouns and nouns) and then the stopword lists were applied to filter out words too general in this context (such as the term *legislation* itself). After this, n-grams from the preprocessed texts were compared against terms in the vocabularies to discover candidate key concepts. Once the results were at hand, the final step was to use a weighting scheme (TF-IDF) to pick only the relevant candidates.

[11] http://tieteentermipankki.fi/wiki/Oikeustiede.
[12] http://eurovoc.europa.eu.
[13] http://finto.fi/koko/fi/.
[14] http://dbpedia.org.

The top scoring[15] candidates were written in RDF format and uploaded to the Semantic Finlex service. The annotations are placed in their own graph[16] using the property *common:autRecSubject* [4]. In addition, the annotations were used to generate tag clouds [7]; [8] to visualize document contents.

5. Publishing Platform and Application Programming Interfaces

The Semantic Finlex service adopts the 5 star deployment scheme suggested by Tim Berners-Lee [9]. The Semantic Computing Research Group has previously proposed an extension to the 5 star scheme by adding two more stars to it [10]. The 6th star is obtained by providing the dataset schemas and documenting them. Semantic Finlex schemas can be downloaded from the service and the data models are documented under the[17] domain. The 7th star is achieved by validating the data against the documented schemas to prevent errors in the published data. Semantic Finlex attempts to obtain the 7th star by applying different means of combing out errors in the data within the data conversion process. The service is powered by the Linked Data Finland[18] publishing platform that along with a variety of different datasets provides tools and services to facilitate publishing and re-using Linked Data.

Following the Linked Data principles all URIs are dereferenceable and support content negotiation by using HTTP 303 redirects. In accordance with the ELI specification, RDF is embedded in the HTML presentations of the legislative documents as RDFa[19] markup. In addition to the converted RDF data, the original XML files are also provided.

To support easier use by programmers without knowledge of SPARQL or RDF, a simplified REST API is provided. This API can be accessed by using the URI patterns and specifying JSON as the preferred content type in the header of the HTTP request. This API also returns its data in the JSON-LD RDF format[20]. Much thought has been given to organize the returned data in a way that is as intuitive as possible and usable also as pure JSON. For example, the affordances provided by JSON-LD *@context* definitions are used to encode language versions of texts in *content_fi* and *content_sv* properties, instead of the user needing to filter the rich literals for their desired language. In addition, URL parameters are provided for retrieving the information pertinent to most common use cases in a stable structure, such as being able to specify which temporal, language, and format versions (of txt and html) of the legislation are required. Finally, a *tree* parameter is provided to build and return the entire subtree of a legislative document without the need to resort to complicated SPARQL queries.

For queries that go beyond fetching information on individual pieces of legislation (such as relational or data analysis queries), a SPARQL endpoint is also available, and a number of sample SPARQL queries are provided to draw inspiration from.

[15] Here, it was decided that the keyword amount is based on document length to have a maximum of 5 to 15 keywords depending on the length of the document. The range was selected based on the analysis of the material [5].

[16] http://data.finlex.fi/annotation/sd.

[17] data.finlex.fi.

[18] http://ldf.fi.

[19] http://www.w3.org/standards/techs/rdfa.

[20] https://json-ld.org/.

6. Application Demonstrators

This Section discusses and presents examples of use cases of the Semantic Finlex data service.

6.1. Data Analysis

Regarding data analysis, sample SPARQL queries were drafted to extract interesting information from the data. These were then fed through to Google Charts for visualization. Information queried was, for example:

- Laws most often referred to from other legislation as well as court decisions.
- Laws that have been changed or amended the most.
- The years in which the above laws were laid.
- The number of EU transpositions by year.
- The members of the supreme court with the most decisions, as well as their tenures.
- The most common topics for supreme court cases by their key concepts.

6.2. Applications

In addition to drafting examples of data analysis, the following application demonstrators were built on top of the Semantic Finlex Linked Data service.

Legal recommender. The HTML representations of the documents are enriched with recommendations to related sources of law similar to [11]. For example, links to relevant EU legislation are queried from the CELLAR SPARQL service[21] by matching their Eurovoc based keywords with the semantic annotations in the Semantic Finlex datasets.

Document-based search. This application can be used to search for similar judgments using an existing document (PDF, image or text) as a search query, for example a court case. The user can choose which search algorithm the application uses to find the documents. Currently the algorithms available are TF-IDF, Doc2Vec and LDA (Latent Dirichlet Analysis).

Search based on text and ontologies. Another text-based search tool with semantic autocompletion [12] has been implemented in connection with the service. The search tool works for both legislation and case law. The search is targeted on different fields in the following order of importance: keywords, document titles, section headings, and texts.

Tag clouds. Tag clouds were used to visualize the contents of the documents. For each statute, a tag cloud was generated using the same process as with the semantic annotations of key concepts described earlier.

Contextual reader. To support users in making sense of the legal terminology in the law and court decision texts themselves, the CORE contextual reader [13] was configured with the legal terminology stored in Semantic Finlex. Figure 5 depicts the user interface, where the user is reading a statute (1) in the Semantic Finlex. CORE enables highlighting each instance of specialist terminology in the documents with a popover providing the definition of that term on a mouse-over (2). In this case, three terminological data sources on the Web are connected to the system, indicated by different colors (3). Further, clicking on any marked mention brings in other laws, decisions, and legal news

[21] http://publications.europa.eu/webapi/rdf/sparql.

articles pertaining to that topic, thus facilitating further semantic browsing and delving more deeply into matters either interesting or unclear (4).

Figure 5. Contextual reader for the Semantic Finlex

Annotation widget. Publishing legislation as Linked Data enables the use of statutes as a reference ontology for linking and integrating heterogeneous datasets that refer to law. The SPARQL endpoint of data.finlex.fi can be utilized as an ontology service [14] for finding relevant statutes or their parts to be used in metadata descriptions. We have prototyped an autocompletion annotation widget that allows the user to search for statutes and fetch their URIs into a cataloging system by typing in a part of the statute name, in the same way as in ONKI [15].

Content widget. As the legislation dataset includes the textual contents of statutes, it can be used to enrich external websites with up-to-date law texts as a web widget [16]. For example, a news article or government announcement informing of a critical change in a statute can be accompanied with the new versions of the relevant parts of the statute. Once the widget is configured with the URI of the desired statute (or its specific part), it will perform a SPARQL query to fetch the text content of the statute, and display it on the web page.

7. Automatic Anonymization and Annotation of Legal Texts

This Section presents our ongoing work in developing tools and services to automatically anonymize and annotate legal texts.

7.1. Automatic Anonymization of Judgments

Due to issues of data protection and privacy, judgments must be anonymized prior to their publication as open data on the Web. Anonymization is the process of removing explicitly or implicitly identifying details of persons and organizations from text. In most Finnish courts, anonymized versions of the judgments are not produced during the court process but the set of documents selected for publication on the Web is anonymized manually as needed. However, anonymization is laborous and costly when done manually. Edita Publishing Ltd. has estimated that it takes approximately 40 minutes to manu-

ally anonymize a single precedent of the Supreme Court. Yet few automatic anonymization tools are currently being used by the Finnish public sector because of the difficulty of evaluating the adequacy of de-identification for different types of data and requirements [17]. Moreover, it is difficult to find a tool able to carry out the anonymization reliably for both Finnish and Swedish language texts.

To facilitate the publication of court cases on the Web and reduce the costs of anonymizing them, we have started developing a configurable general-purpose anonymization service for Finnish and Swedish texts. The service carries out the anonymization process automatically by performing pseudonymization, that is replacing all identifiable information appearing in a text with neutral identifiers. To take into account the different requirements of different actors we make the service configurable so its users can choose what types of identifiable information need to be obfuscated. Taking into account the possibility of error in the automatic anonymization process, the service allows the user to make revisions on the anonymized version using a web-based editor before exporting the final version.

The anonymization service consists of two separate software components, namely a web service and a user interface. The web service takes text as input, finds the named entities in the text and produces as output the same text annotated with special tags that mark the occurrences of the named entities. The special tags contain additional metadata about the occurrences, such as an entity-specific identifier, a category (person, place, organization etc.), and grammatical case. The identifier and category are required so that the occurrences can be correctly transformed into their correct pseudonymized forms such as ‘person A’ or ‘company B’. The information about the grammatical case is needed so that the occurrences of named entities in the text can be replaced with their pseudonyms in correctly inflected forms. To locate the different types of named entities in the text, natural language processing tools such as part-of-speech taggers and different named entity recognition tools are utilized. The recall of the tools is more crucial than the precision [18], as it is more important to hide all critical information than to have some non-critical information incorrectly hidden.

The user interface, shown in Figure 6 is a web-based editor that allows the user to perform the whole anonymization workflow for a document. The document can be either in plain text, XML or HTML format. The anonymization workflow starts by importing a document into the editor. The document is first processed by the web service finding a set of named entity candidates according to a list of preconfigured categories. The document text with the occurrences of named entities highlighted is then shown in the left column of the application view and a list of the named entity candidates along with their pseudonyms is shown in the right column. The user is then able to modify the named entity candidates and the occurrences. After editing is finished, an anonymized version of the document can be exported with all of the occurrences of selected named entities substituted with pseudonyms.

The performance of the service will be compared with the estimates of Edita Publishing Ltd. The success of the tool depends on the applicability of the user interface in addition to the precision and recall of the language technology tools. Therefore, we will incorporate the users and carry out usability tests already in an early phase of the project.

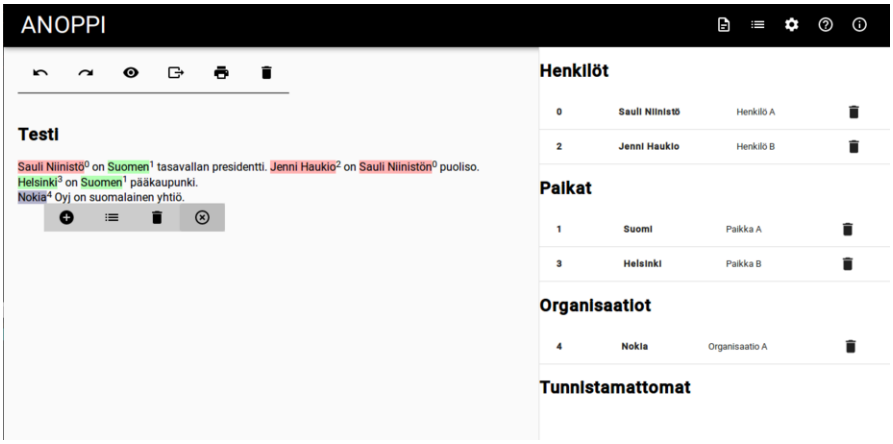

Figure 6. User interface of the anonymization service

7.2. Automatic Annotation of Legal Texts

The same approach used in the automation of the anonymization process can also be applied to the task of automatic or semi-automatic annotation of texts. The difference is that instead of obfuscating the occurrences of named entities found in the text they are linked to existing knowledge bases by applying Linked Data standards. In the context of legislation and case law these entities could be, for instance, key concepts, legal citations, references to parliamentary works or names of judges. Therefore, we plan to use the same user interface and natural language processing tools for the annotation task as for the anonymization, but instead of generating pseudonyms for the named entities, a list of links is provided from which the user can select the correct one. The automatically linked annotations can then serve, for instance, as a basis for the contextual reader application.

8. Related Work and Discussion

Similar efforts to publish legislation and case law as Linked Open Data have been conducted in various countries. The main inspiration for our work was the MetaLex Document Server[22] [19], that provides regularly updated Dutch legislation as Linked Open Data utilizing the CEN MetaLex XML and ontology standards. Another known example of a MetaLex based legal Linked Data service is legislation.gov.uk[23] that hosts UK legislation in local XML formats together with RDF metadata based on the MetaLex ontology. There is also an implementation of a legal Linked Data service in Greece, named Nomothesia[24] [20], that uses both MetaLex and ELI ontologies and ELI-compliant URIs.

Various other ELI implementations and prototypes have also been implemented, usually by resolving ELI-compliant URIs and rendering ELI metadata to existing legal in-

[22] http://doc.metalex.eu.
[23] http://legislation.gov.uk.
[24] http://legislation.di.uoa.gr.

formation portals such as in Luxembourg [25], France [26], and Norway [27]. Many countries already produce ECLI-compliant case law documents to be indexed by the ECLI search engine[28]. Semantic Finlex aims to widen focus by providing both legislation and case law as Linked Open Data through simple Linked Data APIs and linking both of the datasets with each other.

In addition, examples of automatic anonymization and annotation methods applied to legal texts [18]; [21]; [22] and in other domains [23]; [24]; [25] already exist. We aim to apply similar methods to Finnish and Swedish language texts and offer the end-users means to easily revise the possibly incorrect automatic annotations.

A lesson learned during the Semantic Finlex project was that the way the legislation is currently drafted in Finland prevents publishing up-to-date consolidated versions automatically without the need for manual editorial work. To produce such documents without the need for costly editorial work or error-prone automated named entity recognition techniques, the legislative XML and RDF standards, such as Akoma Ntoso [26], should be applied as early as in the legislative and judicial processes where the documents are drafted.

Another issue is that due to copyrights caused by editorial work carried out by the publishing company, we have had to publish the consolidated legislation under a license that restricts its commercial use. The simplest method to circumvent the copyright issues altogether would be to eliminate the need for consolidation by changing the legislative process so that new versions of complete statutes would be published as official versions instead of amendments comprising individual sections.

The new version of the Semantic Finlex service was released on March 10, 2016, and has been in use since. It has been, for example, used in a public hackathon organized by University of Helsinki Legal Tech Lab[29] in October 2017. The development of the service is carried on by further developing the existing application demostrators and tools for automatic anonymization, annotation and validation of the data.

Acknowledgements

The work is a joint effort between Aalto University, the Ministry of Justice, the Ministry of Finance, Edita Publishing Ltd., Business Finland, and Viestintäalan tutkimussäätiö. We thank Risto Talo, Jari Linhala, and Sari Korhonen of Edita Publishing Ltd. for collaboration. The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

[1] Hietanen, A. (2009). Free Access to Legislation in Finland: Principles, Practices and Prospects. In Peruginelli, G. & Ragona, M. (Eds.). *Law Via the Internet. Free Access - Quality of Information - Effectiveness of Rights*. European Press Academic Publishing.

[25] http://legilux.public.lu/editorial/eli.
[26] http://www.eli.fr/en/constructionURI.html.
[27] http://lovdata.no/eli.
[28] https://e-justice.europa.eu/content_ecli_search_engine-430-en.do.
[29] https://www.helsinki.fi/en/networks/legal-tech-lab.

[2] Frosterus, M., Tuominen, J. & Hyvönen, E. (2014). *Facilitating Re-use of Legal Data in Applications. Finnish Law as a Linked Open Data Service*. In Hoekstra, R. (Ed.). *Legal Knowledge and Information Systems. Proceedings of the 27th Jurix Conference*. IOS Press, 115-124.

[3] Lindholm, J. & Derlén, M. (2015). Festina lente – Europarättens genomslag i svensk rättspraxis 1995–2015. *Europarättslig tidskrift*, (1), 151-177.

[4] Tamper, M. et al. (2017). AATOS – A Configurable Tool for Automatic Annotation. In Gracia, J., Bond, F., McCrae, J., Buitelaar, P., Chiarcos, C. & Hellmann, S. (Eds.). *Language, Data, and Knowledge*. LDK 2017. Lecture Notes in Artificial Intelligence. Springer, 276-289.

[5] Tamper, M. (2016). *Extraction of Entities and Concepts from Finnish Texts*. Master's Thesis, Aalto University, School of Science, Degree Programme in Computer Science and Engineering.

[6] Mäkelä, E. (2016). LAS: An Integrated Language Analysis Tool for Multiple Languages. *The Journal of Open Source Software, 1*(6), http://dx.doi.org/10.21105/joss.00035.

[7] Kuo, B. Y., Hentrich, T., Good, B. M. & Wilkinson, M. D. (2007). Tag Clouds for Summarizing Web Search Results. In *Proceedings of the 16th International Conference on World Wide Web*. ACM, 1203-1204.

[8] Hearst, M. A. & Rosner, D. (2008). Tag Clouds: Data Analysis Tool Or Social Signaller? In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*. IEEE, 160-160.

[9] Berners-Lee, T. (2006). *Design Issues: Linked Data*, http://www.w3.org/DesignIssues/LinkedData.

[10] Hyvönen, E., Tuominen, J., Alonen, M. & Mäkelä, E. (2014). Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets. In *Proceedings of ESWC 2014 Demo and Poster Papers*. Springer.

[11] Winkels, R., Boer, A., Vredebregt, B. & van Someren, A. (2014). Towards a Legal Recommender System. In Hoekstra, R. (Ed.). *Legal Knowledge and Information Systems. Proceedings of the 27th Jurix Conference*. IOS Press, 169-178.

[12] Hyvönen, E. & Mäkelä, E. (2006). Semantic Autocompletion. In *Asian Semantic Web Conference*. Springer, 739-751.

[13] Mäkelä, E., Lindquist, T. & Hyvönen, E. (2016). CORE - A Contextual Reader Based on Linked Data. In *Proceedings of Digital Humanities*, 267-269, http://dh2016.adho.org/abstracts/4.

[14] d'Aquin, M. & Noy, N. F. (2012). Where to Publish and Find Ontologies? A Survey of Ontology Libraries. *Journal of Web Semantics, 11*, 96-111.

[15] Tuominen, J., Frosterus, M., Viljanen, K. & Hyvönen, E. (2009). ONKI SKOS Server for Publishing and Utilizing SKOS Vocabularies and Ontologies As Services. In *European Semantic Web Conference*. Springer, 768-780.

[16] Mäkelä, E., Viljanen, K., Alm, O., Tuominen, J. et al. (2007). Enabling the Semantic Web with Ready-to-use Web Widgets. In *Proceedings of the 1st International Conference on Industrial Results of Semantic Technologies*, 293. CEUR-WS.org, 56-69.

[17] Bäck, A. & Keränen, J. (2017). Anonymisointipalvelut. Tarve ja toteutusvaihtoehdot, http://urn.fi/URN:ISBN:978-952-243-503-3. Liikenne- ja viestintäministeriön julkaisuja 7/2017.

[18] Povlsen, C., Jongejan, B., Hansen, D. H. & Simonsen, B. K. (2016). Anonymization of Court Orders. In *11th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, 1-4.

[19] Hoekstra, R. (2011). The MetaLex Document Server. In *International Semantic Web Conference*. Springer, 128-143.

[20] Chalkidis, I., Nikolaou, C., Soursos, P. & Koubarakis, M. (2017). Modeling and Querying Greek Legislation Using Semantic Web Technologies. In *European Semantic Web Conference*. Springer, 591-606.

[21] Lesmo, L., Mazzei, A. & Radicioni, D. P. (2009). Extracting Semantic Annotations from Legal Texts. *Hypertext*, 167-172.

[22] Bartolini, R., Lenci, A., Montemagni, S., Pirrelli, V. & Soria, C. (2004). Automatic Classification and Analysis of Provisions in Italian Legal Texts: A Case Study. In *OTM Confederated International Conferences On the Move to Meaningful Internet Systems*. Springer, 593-604.

[23] Szarvas, G., Farkas, R. & Busa-Fekete, R. (2007). State-of-the-art Anonymization of Medical Records Using an Iterative Machine Learning Framework. *Journal of the American Medical Informatics Association, 14*(5), 574-580.

[24] Kleinberg, B. & Mozes, M. (2017). Web-based Text Anonymization with Node.js: Introducing NETANOS (Named entity-based Text Anonymization for Open Science). *The Journal of Open Source Software, 2*(293).

[25] Mendes, P. N., Jakob, M., García-Silva, A. & Bizer, C. (2011). DBpedia Spotlight: Shedding Light on the Web of Documents. In *Proceedings of the 7th International Conference on Semantic Systems*. ACM, 1-8.

[26] Palmirani, M. (2011). Legislative Change Management with Akoma-Ntoso. In Sartor, G., Palmirani, M., Francesconi, E. & Biasiotti, M. A. (Eds.). *Legislative XML for the Semantic Web: Principles, Models, Standards for Document Management*. Springer.