Facilitating Re-use of Legal Data in Applications—Finnish Law as a Linked Open Data Service

Matias FROSTERUS a,1, Jouni TUOMINEN a and Eero HYVÖNEN a

a Semantic Computing Research Group (SeCo)
Aalto University and University of Helsinki

Abstract Juridical information is important to organizations and individuals alike and is needed in all walks of life. The Finnish government has therefore published Finnish law and related juridical documents on the Web as a service called Finlex. However, even if the documents there are openly available for humans to read, the underlying data has not been open, is based on a traditional XML schema, and does not conform to new semantic metadata standards and Linked Data principles. As a result, the data is difficult to re-use in applications, the datasets are not interoperable with each other, are difficult to link to external data sources, and lots of manual work is needed in producing and using the data. To mitigate these problems, this paper presents Semantic Finlex, the first attempt at publishing Finnish law as a Linked Open Data service, with an analysis and examples of benefits and challenges encountered when applying the technology.

Keywords. Linked Open Data, Legal knowledge publication, Web Services

1. Why Juridical Linked Data?

Publishing and using juridical information is challenging in many ways [14,4]: It is produced by different parties, such as parliaments, ministries, governmental bureaus, various kinds of courts, and research organizations. The content is heterogeneous and produced using differing tools, data formats, and practices. The links between documents are often informal and/or not made explicit. The law in general is a dynamic, changing entity. For example, it is important to be able to refer to different versions of a law at different points in time. These challenges can be addressed through the use of Linked Data technology [5]. Furthermore, in this way linked juridical data can be published not only as data but in the form of ready-to-use services for human end-users and machines.

This paper argues that linked data publication is useful for the following user groups:

1) Content producers. Shared content and metadata formats facilitate the publication and aggregation of contents.

2) Application developers. Availability of linked open data provided with standard application interfaces enables efficient building of more “intelligent” services.

1Corresponding Author: Matias Frosterus, Aalto University, Department of Media Technology, P.O. Box 15500, FI-00076 AALTO; E-mail: matias.frosterus@aalto.fi.
3) Data analysts. Data publication can be accompanied with analytic and visualization tools that help to analyze and interpret the data, e.g., to investigate the mutual dependencies between statutes and the temporal evolution of law.

This paper presents a case study of publishing the Finnish legislation and related documents, such as court cases, as a Linked Open Data service for applications to use. The new service, “Semantic Finlex” is based on the extensive Finlex Data Bank\(^2\) of the Finnish Ministry of Justice. Our goal is that based on the new service, a next generation of more intelligent and useful public and commercial legal systems can be built in a more cost efficient way. We also envision, that by using Linked Data the content production of legal documents and their metadata can be harmonized and made more easily based on established standards, tools, and practices.

In the following, we first discuss how legal documents are produced in Finland. After this, metadata modeling of legal documents in Finlex and juridical domain ontology development is in focus, followed by a presentation of the actual Linked Open Data service now available on the Web. Finally, addressing the needs of the three user groups above via application demonstrators is discussed.

2. Content Production of Juridical Information

The current state of publishing juridical information in Finland is depicted in Figure 1. The Parliament of Finland enacts acts and publishes them as text documents. Based on the acts, ministries and governmental bureaus may issue decrees and regulations. Committee reports, government bills and other preparatory work by governmental parties are produced in various formats. Different levels of courts publish their decisions on cases on their websites. These documents are written for human consumption and every publisher typically has its own preferences on the publishing format (e.g., Word or PDF documents, HTML, XML). If the documents contain metadata, they are produced using a variety of schemas and vocabularies.

The Ministry of Justice publishes parts of this document set in the Finlex Data Bank, a free public service for citizens and organizations. The service contains the Finnish legislation, decisions of courts, international treaties, government bills, and secondary legislation, such as regulations of state authorities. Not all public juridical information is available on the web. For example, one can read the decisions of district courts only by visiting a corresponding court office or ordering them by phone or email.

There are also two major commercial Finnish juridical information portals, Edilex\(^3\) and Suomen Laki\(^4\) (Finnish Law). These portals are aimed for legal experts, such as lawyers and state authorities. The portals contain information that is aggregated from publicly available juridical documents. Journalists of Edilex and Suomen Laki enrich this data by providing additional information and linking it to related contents. Due to the amount of different formats and schemas used, simple harvesting of the data and transforming it into common formats is time-consuming and at least partly manual work.

Our vision of publishing juridical information as Linked Open Data is presented in Figure 2. The goal is to build a harmonized publication process and the tools and services

\(^2\)http://www.finlex.fi/en/
\(^3\)http://www.edilex.fi/
\(^4\)http://www.suomenlaki.com/
for producing and utilizing juridical information as linked data. Standardized formats and schemas allow for the easy collation of data from different sources and linked data techniques allow for common interfaces to be used to access the data. Semantic Finlex is used to harmonize the published content that can be accessed either by using a public portal or by commercial actors to build more specialized and targeted services. Finally, other services and information portals on the web can link to the materials of Semantic Finlex and use widgets to get current legislative data attached to their articles.

3. Modeling Statutes and Court Decisions in RDF

Semantic Finlex contains metadata about the statutes and the cases, i.e., court decisions, the two most used datasets in the Finnish legal domain and Finlex. This section presents metadata schemas\textsuperscript{5} for these datasets, based on the Finlex XML schema, as well as do-

main ontologies used in filling in the values of metadata fields. The core RDFS classes of the model are Statute and CourtDecision.

**Statutes.** Statutes are modelled in the spirit of the FRBRoo model [13], similarly as in the work of Hoekstra [5]. We use a two-level approach by representing statutes on 1) a general level (statute as such), and 2) a version level (statute issued at a certain time). The idea is to give all the general and version level instances dereferenceable URIs so that the model supports diverse ways of linking to legislation. For example, the processes of a company are regulated by some statutes (generally), but a court decision is issued under a statute version that was valid at the time of the court hearing. However, in the first version of Semantic Finlex only the current version of law is considered, and this ambiguity is not a problem.

Statutes consist of smaller items, such as parts, chapters, sections and paragraphs. Statutes and their items are given URIs to allow linking to them inside the legislation data and from other datasets. Statutes may be linked to other statutes (or items) indicating respectively a statute 1) basing on, 2) amending, 3) repealing, 4) entering into force by, or 5) referring to another statute. The subject topic of a statute (item) is expressed with references to concepts in a domain ontology that was created based on legislation related terminology in use in Finland. Similarly, statutes may be classified by linking them to categories of Finnish law classification systems. Finnish law lacks an official categorization but the companies Edita Publishing and Talentum, who publish the Edilex and Suomen Laki services, respectively, have built categories for the statutes based on topic. However, these categorizations are not identical or mapped to one another.

The properties for describing a statute are presented in Table 1. They apply to the version level of a statute because the majority of the statute content is version-specific, whereas the properties relevant on the general level are limited to the statute’s id and references to specific versions of the statute. Individual statute items contain text and structural elements, such as headings and preambles. The statute metadata model reflects the contents and structures of the pilot datasets we have received from the Ministry of Justice, Edilex Inc, and Talentum corp. However, the schema can be easily extended, e.g., with references to preparatory work documents, when new data is incorporated into the system.

<table>
<thead>
<tr>
<th>Label/description</th>
<th>Identifier</th>
<th>Cardinality</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statute name</td>
<td>:title</td>
<td>1</td>
<td>Literal</td>
</tr>
<tr>
<td>Statute identifier</td>
<td>:id</td>
<td>1</td>
<td>Literal</td>
</tr>
<tr>
<td>Date of adoption</td>
<td>:adoptionDate</td>
<td>1</td>
<td>Literal</td>
</tr>
<tr>
<td>Date of entry into force</td>
<td>:entryIntoForceDate</td>
<td>1</td>
<td>Literal</td>
</tr>
<tr>
<td>Has part (statute item)</td>
<td>:hasPart</td>
<td>0..n</td>
<td>Statute item instance</td>
</tr>
<tr>
<td>Enacting clause</td>
<td>:enactingClause</td>
<td>1</td>
<td>Enacting clause instance</td>
</tr>
<tr>
<td>Sub-heading</td>
<td>:subHeading</td>
<td>0..n</td>
<td>Sub-heading instance</td>
</tr>
<tr>
<td>Subject topic</td>
<td>:subject</td>
<td>0..n</td>
<td>Concept</td>
</tr>
<tr>
<td>Classification category</td>
<td>:category</td>
<td>0..n</td>
<td>Classification category</td>
</tr>
<tr>
<td>Bases on statute</td>
<td>:basesOn</td>
<td>0..n</td>
<td>Statute instance</td>
</tr>
<tr>
<td>Amends statute (item)</td>
<td>:amends</td>
<td>0..n</td>
<td>Statute (item) instance</td>
</tr>
<tr>
<td>Repeals statute (item)</td>
<td>:repeals</td>
<td>0..n</td>
<td>Statute (item) instance</td>
</tr>
<tr>
<td>Entry into force by statute</td>
<td>:entryIntoForceBy</td>
<td>1</td>
<td>Statute instance</td>
</tr>
<tr>
<td>Refers to statute (item)</td>
<td>:refersTo</td>
<td>0..n</td>
<td>Statute (item) instance</td>
</tr>
</tbody>
</table>

Table 1. The properties for describing a statute.

**Court Decisions.** Decisions of the supreme court and the court of appeal are published in Finlex while the various district courts’ decisions are not openly available. The
cases refer to laws but formal, machine-understandable links are provided only occasionally and the references are not easy to parse automatically since they can be made with abbreviations or by using the names of individual statutes.

A court decision consists of a summary text and text paragraphs with optional headings. A court decision may be linked to statutes it is issued under, and to the previous court decisions given regarding the case. The subject topic of a decision may be given by linking the decision to a concept in a legal ontology. The court that issued the decision and the individual decision-makers involved are expressed with references to their URIs in court and actor ontologies. The properties of a court decision are presented in Table 2.

<table>
<thead>
<tr>
<th>Label/description</th>
<th>Identifier</th>
<th>Cardinality</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision name</td>
<td>:title</td>
<td>1</td>
<td>Literal</td>
</tr>
<tr>
<td>Decision identifier</td>
<td>:id</td>
<td>1</td>
<td>Literal</td>
</tr>
<tr>
<td>Date of issue</td>
<td>:dateOfIssue</td>
<td>1</td>
<td>Literal</td>
</tr>
<tr>
<td>Summary</td>
<td>:summary</td>
<td>1</td>
<td>Literal</td>
</tr>
<tr>
<td>Has paragraph</td>
<td>:hasParagraph</td>
<td>0..n</td>
<td>Paragraph instance</td>
</tr>
<tr>
<td>Subject topic</td>
<td>:subject</td>
<td>0..n</td>
<td>Concept</td>
</tr>
<tr>
<td>Issued by court</td>
<td>:court</td>
<td>1</td>
<td>Court instance</td>
</tr>
<tr>
<td>Issued under statute (item)</td>
<td>:issuedUnder</td>
<td>0..n</td>
<td>Statute (item) instance</td>
</tr>
<tr>
<td>Previous court decision</td>
<td>:previousDecision</td>
<td>0..n</td>
<td>Court decision instance</td>
</tr>
<tr>
<td>Decision-maker</td>
<td>:decisionmaker</td>
<td>1..n</td>
<td>Person instance</td>
</tr>
</tbody>
</table>

Table 2. The properties for describing a court decision.

Linked Domain Ontologies. Annotations to Finnish juridical documents have been provided by different bureaus of judicial administration and also by the publishers Talen- tum and Edita Publishing. The Finnish Ministry of Justice has collected 26 different vocabularies used by the judicial administration and combined them into a single vocabulary called Asseri of almost 6000 terms. Talentum’s Suomen Laki Index of Terms is comprised of some 3800 terms and Edilex Vocabulary of Edita Publishing contains nearly 4000 terms.

The overlaps between these vocabularies as well as the Finnish General Upper Ontology YSO [6] are depicted in Table 3. The numbers in parentheses on the diagonals show the size of the vocabulary and the percentage values show how much overlap there is. For example, the second column on the first row shows that Suomen Laki vocabulary covers six percent of YSO. The calculation of the overlap was done simply by string comparison. The amount of hierarchical or other structure in the vocabularies varied widely so it was concluded that using it in the overlap detection would not lead to more accurate results. From the percentages presented, we can see that the legal vocabularies have considerable overlap with each other as well as with the general ontology but that each vocabulary also features many terms not present in others.

<table>
<thead>
<tr>
<th></th>
<th>YSO</th>
<th>Suomen Laki</th>
<th>Edilex</th>
<th>Asseri</th>
</tr>
</thead>
<tbody>
<tr>
<td>YSO</td>
<td>(25893)</td>
<td>6%</td>
<td>6%</td>
<td>12%</td>
</tr>
<tr>
<td>Suomen Laki</td>
<td>43%</td>
<td>(3778)</td>
<td>45%</td>
<td>49%</td>
</tr>
<tr>
<td>Edilex</td>
<td>41%</td>
<td>(3974)</td>
<td>45%</td>
<td></td>
</tr>
<tr>
<td>Asseri</td>
<td>51%</td>
<td>31%</td>
<td>31%</td>
<td>(5862)</td>
</tr>
</tbody>
</table>

Table 3. Number of overlapping concepts in the three legal vocabularies and the General Finnish Upper Ontology YSO.
Using ontological concepts with URIs instead of string-based thesaurus terms is useful in many ways [1]. Unique identifiers allow references to concepts to remain constant if the preferred labels change. URIs allow additional information to be accessed easily when using the dereferencing mechanism of Linked Data. Concept hierarchies make reasoning possible. Using language independent concepts is especially pertinent in Finland due to the two official languages, Finnish and Swedish, and therefore all legislation and much of the accompanying material needs to be available in both languages.

We have begun the ontology work by harmonizing different in-use thesauri in the field of the Finnish legislation and by building an ontology of legal concepts. The three vocabularies (Asseri, Suomen Laki, Edilex) were mapped to each other using a simple label-matching procedure. The resulting vocabulary contains ca. 9,000 different terms. Unfortunately, all of them are available only in Finnish. The terms were then mapped to the larger KOKO ontology cloud [6] (ca. 45,000 concepts) that is collated by the National Library of Finland and caters to the needs of various domains. The KOKO cloud consists of a general upper ontology and domain ontologies from a wide spectrum of life. KOKO acts as a conceptual hub for digital collections of different kinds. Still, human work is needed for validating the mappings and refining the semantic relations of the concepts to ensure, e.g., the integrity of the transitive subclass-of relations.

4. The Linked Open Data Service

We have published the pilot datasets of the Finnish legal system in the "Linked Data Finland" platform LDF.fi\(^6\) [7]. The datasets are based on the data dumps we received from the Finlex Data Bank in 2012: 1) legislation valid at that time: 2413 statutes, 2) supreme court yearbooks: 11904 decisions (1926–2012), and 3) supreme administrative court yearbooks and summaries of other cases: 1490 decisions (1944–2012). LDF.fi extends the 5-star model of publishing Linked Data\(^7\) into a 7-star model by emphasizing the importance of providing explicit metadata schemas for the data (the 6th star) and validating the data against the schemas (the 7th star). These additional requirements for high-quality linked data allow people to understand and re-use the data more easily, and to trust in the quality of the data. The public availability of schemas and validation of datasets according to them is important in the field of legal information for ensuring the quality and interoperability of the datasets, as there is a variety of content types and producers involved, raising the probability of inconsistencies in the data publication.

LDF.fi automates the Linked Data publishing process by generating a rich set of data services based on the metadata description of the dataset, which uses an extended version of the SPARQL Service Description\(^8\) and the VoID\(^9\) vocabularies. The data services include a SPARQL endpoint, dereferenceable URIs, dataset download, browsing, editing, validation, visualization, and schema documentation and download.

The LDF.fi service uses the Fuseki SPARQL server\(^10\) for storing and accessing the RDF data, and Varnish\(^11\) for URL redirects for dereferenceable URIs and other data ser-

\(^6\)http://www.ldf.fi/dataset/finlex
\(^7\)http://www.w3.org/DesignIssues/LinkedData.html
\(^8\)http://www.w3.org/TR/sparql11-service-description/
\(^9\)http://rdfs.org/ns/void
\(^10\)http://jena.apache.org/documentation/serving_data/
\(^11\)https://www.varnish-cache.org/
The browsing and editing of a dataset is provided with the SPARQL SAHA system. The vocab.at service\(^ {12}\) is used for analyzing a dataset in terms of schema (vocabulary) usage and reporting quality issues, such as the usage of undefined properties and unresolvable namespaces. Statistical description (extended VoID) of a dataset is generated using the Aether tool\(^ {13}\). The statistics can then be viewed, e.g., for getting an overview of the dataset or comparing two versions of it. A dataset can also be visualized with the Sgvizler tool\(^ {14}\) or dataset-specific tools if provided. Schema documentation is provided by LODE\(^ {15}\) in human-readable HTML form.

The legal ontologies and statutes are published also in the ONKI ontology service\(^ {16}\), providing not only a browsing facility but a selector widget with autocompletion for annotation, and APIs for machine use \(^ {16}\). Each statute and its version has a URI that can be found and fetched as an unambiguous reference from ONKI, to be used, e.g., for annotating court decisions with statutes they are issued under.

5. Evaluating and Demonstrating Results

To test the applicability of our approach to publishing legal linked data, we are building proof-of-concept demonstrators for addressing the needs of the three user groups discussed in Section 1.

**Content producers.** The RDF-based metadata models for legislation and case law allows the heterogeneous network of content producers to publish their data in an interoperable way. Different datasets can be linked together either directly or via the use of shared legal ontologies and the ONKI ontology service. The linked data publishing platform LDF.fi is an efficient way of publishing the datasets with rich data services. The SAHA tool can be used for editing and browsing the datasets. We have published the pilot datasets of Semantic Finlex utilizing these models and services. We have tested the linking of an external dataset to the legislation data with a legal news dataset published by Edilex. The original dataset was published on the web with hyperlinks from news items to relevant statutes. We converted the XML dump of the news data into RDF format and linked the articles to the statute URIs of the Semantic Finlex. The resulting dataset including the titles, publication dates, source URLs, and subject annotations of the news items is published in the LDF.fi service\(^ {17}\), too.

**Application developers.** The legal information can be utilized as machine-processable RDF data via Linked Data best practices, including URI dereferencing, SPARQL endpoint and data dump download. In addition to these baseline data services legislation data can also be utilized as widgets [12] that can be integrated into web pages for providing up-to-date legislation information. For example, a government agency can incorporate legislation relevant to their field into their web pages. An example demonstrator widget is depicted in Figure 3. Here, the widget is used to accompany a news item concerning the smoke-free transition of Finnish restaurants with the relevant part of the Tobacco

---

12\url{http://vocab.at}
13\url{http://demo.seco.tkk.fi/aether/}
14\url{http://dev.data2000.no/sgvizler/}
15\url{http://www.essepuntato.it/lode}
16\url{http://onki.fi}
17\url{http://www.ldf.fi/dataset/media}
Act. The widget is implemented as a JavaScript component that fetches the legislation text with a SPARQL query using the URI of the relevant part of legislation.

**Figure 3.** Widget is used to accompany a news item with relevant legislation text. Original web page courtesy of Helsingin Sanomat.

**Data analysts.** The publication of legal data in harmonized format allows efficient processing of it for conducting analysis with, e.g., visualization toolkits. As part of the LDF.fi service the Semantic Finlex dataset can be explored\(^\text{18}\) with the SPARQL-based visualization tool Sgvizler. We have also visualized the data with the visualization playground VISU\(^\text{19}\), e.g., to show the statutes that are the most referred to by other statutes.

In addition to these user groups directly utilizing the data publication services, we have made some experiments regarding end-user services, e.g., for citizens. New semantic search and recommendation facilities for Semantic Finlex are to be tested using our SAHA-HAKO tool\(^\text{20}\)[8], a metadata editor which incorporates a faceted search engine on top of SPARQL. In this case, useful facets include the statutes and their items, time, subject matter, as well as the law classification systems in use. The recommendation system would allow, e.g., studying a given court case with links to similar cases that reference to the same statutes. This can be further refined by noting the cases where the statute has changed, so that these can be viewed separately.

6. Related Work and Discussion

Our work can be seen as a direct continuation of the global trend of publishing legal information as seen in, e.g., France\(^\text{21}\), the UK\(^\text{22}\), Italy\(^\text{23}\), and Netherlands\(^\text{24}\). In our ap-

---

18\[http://www.ldf.fi/visualization/finlex\]
19\[http://demo.seco.tkk.fi/visu/#/?sparqlEndpoint=http%3A%2F%2Fldf.fi%2Ffinlex%2Fsparql\]
20\[http://code.google.com/p/saha/\]
21\[http://www.legifrance.gouv.fr/\]
22\[http://www.legislation.gov.uk/\]
23\[http://www.normattiva.it/\]
24\[http://wetten.nl/ and http://doc.metalex.eu/\]
proach, we hope to emphasize the re-usability of the legal data when published in our 7-star model of linked open data.

Hoekstra [5] presented an approach for converting Dutch regulation documents from legacy XML format to CEN Metalex[^25], an open XML interchange format for legal and legislative resources, and to RDF format, published as linked data. The work involves the generation and publication of regulation metadata and version information, which is not available in the original sources. Our work follows the same general principles, but we also try to extend the model to court decisions and the Finnish domain.

Eunomos is a legal knowledge management system that employs ontologies [3] but it is based on XML and does not use linked data methods thus making for a less re-usable system. L`egilocal [10] provides a multifaceted approach to accessing multiple types of documents with different structures while taking into account both temporal aspects and the semantic content of the data. The system uses ontologies and supports SPARQL queries. Compared with them, LDF.fi aims to cater especially to practical data services and the re-use of the data in various mash-up services.

The current, very heterogeneous publication of legal material in Finland is easy to improve upon. Varied schemas and formats result in a lot of extra work when collating data. Following the open linked data principles provides easy access to the material and allows for its use in various contexts and applications. Since legal information is often needed in a very precise way, the 7-star model as presented in this paper is even more important than in some other domains.

Because the datasets of the Finnish legislation are annotated at rather high level and the links between different datasets are not explicit in many cases, the utility of the datasets could be enhanced by their semantic enrichment. (Semi)automatic subject annotation [15] of the statutes and identification of statute references in the court decisions [9] seems like a promising approach. In the pilot, we were not able to utilize the FRBR-inspired model fully since we had access only to the current version of legislation. Similarly, in order to realize the full potential of the linked data publication, we would like to publish and link other datasets, such as preparatory works, which are often critical when interpreting the laws. To continue our work, we are planning on utilizing European Legislation Identifiers (ELI)[^26] and European Case Law Identifiers (ECLI)[^27] for compatibility with the EU practices similarly to [2]. This would enable building automatic links between cases in different European countries referring to the same EU legislation.

One interesting topic to tackle is the societal impact of releasing juridical data as linked data. The highly structural nature of the data allows for more powerful and sophisticated analysis of the data possibly leading to new challenges regarding, e.g., privacy.

Based on the hands-on experiments presented in this paper we are confident, even if more formal evaluations are still missing, that the Linked Data approach provides a quite promising basis for publishing, aggregating, and re-using legal data in web applications. However, this would require lots of collaboration and standardization work between the various organizations involved in producing and using legal documents.

[^25]: http://www.metalex.eu/
[^27]: https://e-justice.europa.eu/content/european_case_law_identifier_ecli-175-en.do
Acknowledgements The work is a collaboration with the Ministry of Justice, the Ministry of Communications, Editia Publishing Ltd, and Talentum Corp. under the national Linked Data Finland project funded mainly by Tekes—the Finnish Funding Agency for Technology and Innovation. We thank Mika Wahlroos for collaboration.

References


http://www.seco.tkk.fi/projects/ldf/