*Legal Knowledge and Information Systems*
*A. Wyner and G. Casini (Eds.)*

*doi:10.3233/978-1-61499-838-9-101*

# On Annotation of the Textual Contents of Scottish Legal Instruments

Adam WYNER [a,1] Fraser GOUGH [c], Francois LEVY [b], Matt LYNCH [c], and Adeline NAZARENKO [b]
[a] *University of Aberdeen, Aberdeen, United Kingdom*
[b] *LIPN, Paris 13 University – Sorbonne Paris Cité & CNRS, Paris, France*
[c] *Parliamentary Counsel Office, Scottish Government, Edinburgh, United Kingdom*

**Abstract.** LegalRuleML is a developing standard for representing the fine-grained semantic contents of legal texts. Such a representation would be highly useful for Semantic Web applications, but deriving formal rules from the textual source is problematic; there is currently little in the way of methodology to systematically transform language to LegalRuleML. To address this, we outline the purposes, processes, and outputs of a pilot study on the annotation of the contents of Scottish legal instruments, using key LegalRuleML elements as annotations. The resulting annotated corpus is assessed in terms of how well it answers the users' queries.

**Keywords.** semantic annotation, legal text processing, markup language, methodology

## 1. Introduction

There is an increasing demand for tools enabling fine-grained semantic access to legal sources, that is, for search tools that go beyond keyword search [2], such as Semantic Web applications to link, search, extract, and draw inferences with respect to the contents of and relations amongst legal rules. This paper presents a pilot study that shows how some of these demands, *e.g.* for search and extraction, can be addressed. However, formalizing the rule information present in legal sources is a complex task that cannot be automated due to the complexities of legal language and information. Nonetheless, some progress can be made to annotate the semantic structure of the source texts as well as to comply with existing documents and legal rule standards to ensure interoperability and linkability. The challenges are twofold: to make annotations that address users' interests; to make the annotation task feasible for legal people and in a form amenable to incremental refinement. In the experiment reported, a small corpus of legal instruments is translated to LegalRuleML, an XML mark-up language for legal rules [1]. By way of evaluation, we used the sample questions provided by the use case partners to query the annotated corpus; the results demonstrate the utility of the approach.

In the following, we present the use case requirements (Sec. 2), the annotations and corpus (Sec. 3), the methodology and tools (Sec. 4) and our preliminary outputs (Sec. 5).

[1] Corresponding Author: azwyner@abdn.ac.uk. We thank the funding from the University of Aberdeen's Impact, Knowledge Exchange, and Commercialisation Award for this 10 week study. This work was also supported by the French National Research Agency (ANR-10-LABX-0083) in the context of the Labex EFL. We also thank the student staff: A. Andonov, A. Faulds, E. Onwa, L. Schelling, R. Stoyanov, and O. Toloch.

## 2. Requirements

We started with the requirements set by the parliamentary counsel of the Scottish Government's Parliamentary Counsel Office, which is working to improve internal legislative drafting and information services and to provide legislative information "as a platform" for a robust ecosystem of legal services. A key part of this effort is to provide a corpus of law in electronically readable form which can be queried.

We were provided with questions to answer:

1. What are all the offences and associated penalties or defences?
2. What prohibitions apply to tobacco products?
3. What obligations have been placed on which entities, *e.g.* shop owners?
4. What permissions are given to Scottish Ministers?
5. Given a provision, what are related overriding or reparation provisions?

Answering such questions requires a substantial semantic analysis of the text. The challenge is to develop a level of analysis and XML representation which satisfies the questions. A sound methodology of annotation is necessary to get a corpus that can be further used as a gold-standard for evaluation and machine learning.

## 3. LegalRuleML and Annotations

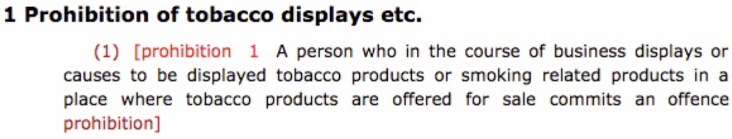
Large corpora of legal texts must be machine-readable [2]. XML standards have been developed for document structure (Akoma Ntoso[2]) and semantic content (LegalRuleML [1]). Complying with such standards allows materials to be amenable to Semantic Web technologies. Yet analysing the semantic content of legal documents in terms of XML is particularly daunting given the nature of linguistic representation; there is a significant gap between the linguistic and formal representations of the law.

LegalRuleML is a proposed OASIS standard for rich XML representation, which has elements to represent legal content. In addition, it adopts a restricted set of XML elements from RuleML, a markup language for predicate logic rules[3]. In order to develop the means to translate from natural language to LegalRuleML, it has been argued that some intermediate annotation language is essential to get a "first draft" of the contents of the legal text as well as to help address linguistic ambiguities and interpretive issues [5]. In this project, we only used a small palette of LegalRuleML elements which associate with text annotations:

- Permission: the bearer is allowed to do something or be in a state.
- Obligation: the bearer is bound to do something or be in a state, for otherwise, the bearer is in violation.
- Prohibition: the bearer is bound not to do something or be in a state, for otherwise, the bearer is in violation.
- Constitutive: a definition.
- Override: an indication that one legal rule takes precedence over another.
- Reparation: an indication of a link between a penalty and a prescriptive norm.
- Penalty: a sanction.

[2] http://www.akomantoso.org/

[3] http://ruleml.org/index.html



**1 Prohibition of tobacco displays etc.**

(1) [prohibition 1 A person who in the course of business displays or causes to be displayed tobacco products or smoking related products in a place where tobacco products are offered for sale commits an offence prohibition]

**Figure 1.** Annotations on text

This small, coarse-grained palette of LegalRuleML elements was useful in addressing some key initial issues. Given an iterative, extensible development process, we can work with other elements in later phases. Similarly and for our purposes here, we do not work with document structure, which would be annotated in Akoma Ntoso, though in future iterations, such information will be important.

While LegalRuleML is explicit, application of the elements to text is not transparent. That is, the list of elements and their definitions are not sufficient for the consistent and accurate application of the annotations to text, nor is there clarification about how to analyse source text into LegalRuleML. Thus, an annotation methodology is required to connect text to LegalRuleML.

## 4. Methodology, Corpus, and Tools

To use LegalRuleML elements for annotation, we "hide" the technical structure of LegalRuleML from legal annotators, whose task was to understand the content. We provided annotators with a simplified set of annotations, where the relevant sentences are bracketed, labelled/typed, and possibly related via indices (see Figure 1). It is important to emphasise that we have "repurposed" LegalRuleML elements as labels/types for text annotation in order to associate text annotations with LegalRuleML representations; we have not thereby created an auxiliary markup language. On the semantic side, we developed guidelines with illustrations of regular and irregular examples to help tackle semantic issues. Adjudication and revision (of annotations and/or the guidelines) were essential.

The project employed four annotators for six weeks; they were students from different disciplines, but with some legal and linguistic training. Each original document was annotated by two legal annotators, who reviewed and commented on one another's work. Three "meta" annotators adjudicated the annotated documents. Once adjudicated, the resulting documents were translated into valid LegalRuleML files by LegalRuleML analysts. The annotators used an annotation manual, which was developed to guide annotations. During the annotation process, comments were added to the document, facilitating and tracking discussion. We reported the main issues and ambiguities in the manual.

For a corpus of texts, we have 10 legal instruments provided by the Scottish Government's Parliamentary Counsel Office (41,859 words, $\sim$ 140 pages)[4]. All bear on Scottish smoking legislation and regulation. The average word count per document is 4185.9, with a maximum word count of 12739 and a minimum of 437. We do not report sentence numbers, for sentence identification in legal text is a difficult, unresolved problem [4].

[4] For a sample of the documents, see http://www.legislation.gov.uk/asp/2016/3, http://www.legislation.gov.uk/asp/2016/14/part/1/chapter/1 or http://www.legislation.gov.uk/ssi/2010/407/made

The workflow was managed on Trello. The (annotated) documents were stored in shared Google Docs directories, which corresponded to the annotation steps. Github served as a code and XML repository. The XML annotated files were transfered to a web site on which they can be queried by XQuery and re-visualised using XSLT.

Some points of disagreement between annotators lead us to revise and clarify the annotation guidelines. Many questions focused on the scope of the annotations and more explicit guidelines have been given, *e.g.* in case of lists and complex sentences. The interpretations of modal verbs, like “may” or “must”, also raised questions as they cannot be automatically matched to one type of prescriptive statement. Examples have been added to draw annotators’ attention on these issues. In some cases, legal annotators lacked the basics of logical reasoning and needed additional explanation (*e.g.* the negation of an obligation is a permission). The annotation of reparations and exceptions appeared to be particularly difficult, probably because of the diversity of possible formulations: the guidelines have been enriched with examples and interpretation tips.

## 5. Results

In this section, we discuss the project outputs, which are:

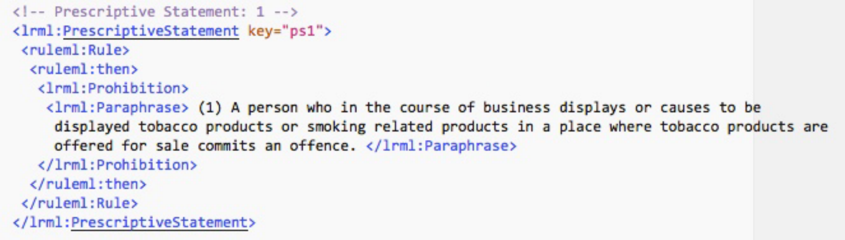
- A very simple annotation language designed for legal annotators and for an automatic transformation into LegalRuleML compliant annotations.
- An annotation manual which provides 1) guidelines for the homogeneous application of legal semantic annotations and 2) instructions on the workflow.
- An annotated corpus and its corresponding LegalRuleML encoding. Presently, 558 statements are annotated.
- A dedicated web application[5], for retrieving the annotated statements based on their types as well as on the keywords or text patterns they contain.

In Figure 1, we have a snippet of source text annotated as a prohibition. Opening and closing brackets indicate the beginning and ending of the annotated text span. A number is introduced to facilitate relating expressions. In Figure 2, we provide the corresponding expression in LegalRuleML. Note that the XML structure requires auxiliary information not found in the source text with annotation, including `PrescriptiveStatement`, a (bodiless) `Rule` with conclusion `then`, a deontic element `Prohibition`, all wrapping the full text as a `Paraphrase`. Note that within `Paraphrase`, we have copied the source text. Thus, our approach maintains the source text for further analysis *in situ*, while wrapping it in valid LegalRuleML. Finally, Figure 3 presents the statement amongst the query results for both “offence” and ”tobacco products” contained within a `PrescriptiveStatement` that is a `Prohibition`.

Most of the questions listed in Section 3 can be answered using the search tool:

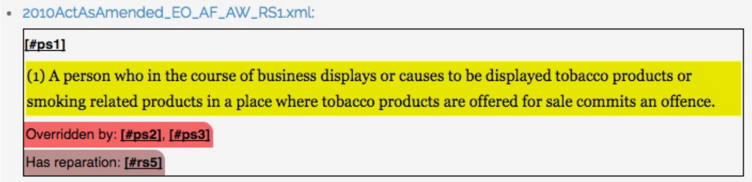
1. All the definitions of offences involve the word “offence”. Searching this word yields 70 statements of different kinds. To focus on definitions, we require also that the statement be a `Prohibition`, which reduces to 26 answers (recall 1, precision .84). Associated defenses are obtained by searching `Permission` elements which contain any of “defence” or “offence” (recall 1, precision .60). In

[5] http://tal.lipn.univ-paris13.fr/LexEx



```
<!-- Prescriptive Statement: 1 -->
<lrml:PrescriptiveStatement key="ps1">
 <ruleml:Rule>
  <ruleml:then>
   <lrml:Prohibition>
    <lrml:Paraphrase> (1) A person who in the course of business displays or causes to be
     displayed tobacco products or smoking related products in a place where tobacco products are
     offered for sale commits an offence. </lrml:Paraphrase>
   </lrml:Prohibition>
  </ruleml:then>
 </ruleml:Rule>
</lrml:PrescriptiveStatement>
```

**Figure 2.** LegalRuleML Representation



- 2010ActAsAmended_EO_AF_AW_RS1.xml:

[#ps1]

(1) A person who in the course of business displays or causes to be displayed tobacco products or smoking related products in a place where tobacco products are offered for sale commits an offence.

Overridden by: [#ps2], [#ps3]

Has reparation: [#rs5]

**Figure 3.** Query Result for “offence” and ”tobacco products” in Prohibition

both cases, all the erroneously recovered statements do not specify the offence but the procedure which applies in case of offence. Last, defenses and reparations are linked to their corresponding offense *via* relations which appear on Fig 3.

2. Enumerating prohibitions which apply to tobacco products is more difficult because of alterantive lexicalisations. A search for “tobacco product[s]” in `Prohibition` elements gets 6 statements. But for “It is an offence for an adult to smoke in a private motor vehicle when there is a child in the vehicle”? terminological knowledge would help. When interpreting “A person who fails to comply with a requirement made under subsection (1) or (2) commits an offence”, one needs to refer to subsections (1) and (2).
3. Obligations placed on shop owners are, for similar reasons, difficult to select. “Shop” appears only once in the texts and “owner” never, “business” being the more usual term, but also “management”, “control”, and “responsible person”.
4. Permissions given to Scottish Ministers are easier to focus on because the title is always literally used. Querying “Scottish Ministers” in `Permission` elements yields 21 statements (precision .952, recall .875). On one side, 1 permission is given to ”a person” ; on the other side, 3 additional permissions are incidentally mentioned in `Obligation` or `Constitutive` statements.
5. As can be seen in Fig 3, related overriding or reparation provisions are mentioned and accessible through a direct link in the display.

## 6. Discussion

Ours is not the first work to attempt the semantic annotation of legal rules, *e.g.* [6]. However, it is the first to tie the annotation effort directly to some well-developed, standardised markup language such as LegalRuleML. In our view, and following [3], the development of a high quality annotator manual which leads a team of annotators to a high level of inter-annotator agreement is an essential task in its own right. Setting up an efficient and simple workflow of annotation is also important if one wants annotators to concentrate on interpretative issues.

Returning to Figures 1-2, our methodology highlights an important issue in formalising source text: annotation requires analyzing the expression. As is apparent, we have taken the "naive" approach of annotating a whole sentence according to key words; that is, (1) is marked as a prohibition given *commits an offence*. Yet, obviously, this is misleading since the contents of the whole annotated text, including *commits an offence*, is not what is prohibited. Rather, what is prohibited is the action *displays or causes to be displayed tobacco products or smoking related products in a place where tobacco products are offered for sale* committed by *a person in the course of business*. What the search tool ought to return is just those prohibitions with respect to their content. Providing such analysis requires some care so as not to distort the meaning of the source expression. Yet, it is such fine-grained analyses that LegalRuleML requires. Our simplified, incremental approach to annotation is but one step towards this more refined result, whilst highlighting problems to address as well as yielding useful results along the way.

Finally, some of the missing results are matters beyond LegalRuleML, *e.g.* lexical semantic relationships amongst terminology. There are interesting interpretive issues concerning linguistic expressions of the annotations, complex expressions, ellipsis, reference, and others. Nonetheless, an advantage of our effort is to draw out a detailed, extensive range of such matters. Thus, there remains significant work ahead.

Now that the annotation guidelines and process have been tested and revised thanks to the adjudication work, a larger annotation experiment can be launched. The quality of the resulting annotated corpora (measured as the inter-annotator agreement) is a key feature, as our ultimate goal is to use it as training data for automating (part of) the annotation process.

## References

[1] Tara Athan, Guido Governatori, Monica Palmirani, Adrian Paschke, and Adam Z. Wyner. Legalruleml: Design principles and foundations. In Wolfgang Faber and Adrian Paschke, editors, *Reasoning Web. Web Logic Rules - 11th Int. Summer School, Berlin, Germany, 2015, Tutorial Lectures*, pages 151–188. Springer, 2015.

[2] Pompeu Casanovas, Monica Palmirani, Silvio Peroni, Tom M. van Engers, and Fabio Vitali. Semantic web for the legal domain: The next step. *Semantic Web*, 7(3):213–227, 2016.

[3] Karën Fort. *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. Wiley-ISTE, July 2016.

[4] Jaromir Savelka and Kevin Ashley. Extracting case law sentences for argumentation about the meaning of statutory terms. In *Proc. of the 3rd Workshop on Argument Mining (ArgMining@ACL 2016)*, Berlin, Germany, 2016.

[5] Adam Wyner, Adeline Nazarenko, and Francois Lévy. Towards a high-level controlled language for legal sources on the semantic web. In Brian Davis, J. Gordon Pace, and Adam Wyner, editors, *Proc. of the 5th Int. Workshop on Controlled Natural Language (CNL2016)*, pages 92–101, Aberdeen, UK, July 2016. Springer.

[6] Adam Wyner and Wim Peters. On rule extraction from regulations. In Katie Atkinson, editor, *Legal Knowledge and Information Systems - JURIX 2011: The Twenty-Fourth Annual Conference*, pages 113–122. IOS Press, 2011.