Why it is necessary for legislators to annotate legislation with meta-data

Rob OPSOMER [a], Geert DE MEYER [b], and Greet VAN EETVELDE [a]

[a] *Ghent University, Dep. of Civil Techniques, Environmental and Spatial Management*

[b] *Flemish Institute for Technological Research*

Abstract. Everyone is supposed to know the law. Since it is not possible for all persons to know all law, persons should be enabled to retrieve all legislation applicable to a certain issue. We argue that this should be done by tagging laws with labels indicating their content. Although such techniques were already introduced years ago, they are still only deployed on a very small scale. The main problem with existing approaches is that the laws are not tagged by the legislators themselves, but by third party legislation managers. This third party tagging is either done manually by legal experts, or automated with the help of artificial intelligence techniques. We argue that both of these approaches are inherently flawed, as automated approaches do not work well, and manual tagging is prohibitively slow, expensive, and error-prone. Therefore, it is necessary that legislation is tagged at creation time, by the same legislators who create it.

Keywords. legislation, legislator, information retrieval, legal text retrieval

1. Introduction

Nemo censetur ignorare legem and *Ignorantia juris non excusat* are two of the best known legal adages. Both have essentially the same meaning: *everyone is supposed to know the law*. It is a fundamental property of all modern legal systems [1]. It is a necessary property: accepting ignorance of law as a defense would make it impossible for courts to function properly [2], as every defendant could plea he did not know the law he was breaching. Since it is such a fundamental property, governments should enable their citizens to fulfill it.

There exists too much legislation, however, to expect every person to know every area of legislation by heart. According to Matthijs, there are about 45000 active laws in Belgium [3]. Even worse, the publishing rate of legislative documents, visualized for Belgium in Figure 1[1], is both very high and ever-increasing, making it practically impossible to stay up to date with all new legislation. Now should it be concluded that one of the most fundamental adages in modern legal systems is unattainable? Not exactly. While it is impossible to be well informed about *all* legislation, this is not a strict necessity in practice. Indeed, more important is that every person knows the legislation *relevant* to him. Of course, to be able to *know* all legislation relevant to a certain issue,

[1] Source: Belgian Federal Government, `http://www.ejustice.just.fgov.be/wet/wet.htm`

persons should first of all be able to *find* all legislation relevant to that issue. Reading *all* laws is obviously not an option. Therefore, tools that enable persons to easily find legislation about a certain topic are required. Thus, the government duty of enabling citizens to know the law boils down to creating good *information retrieval* tools.

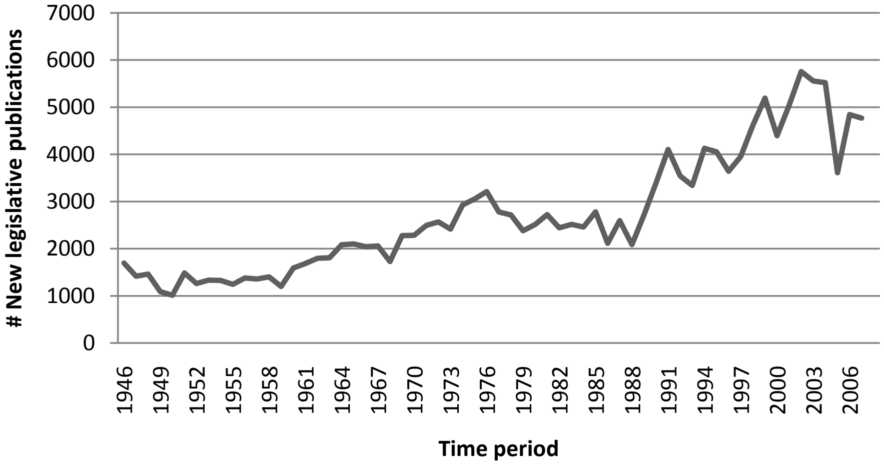

Figure 1. Number of new legislative publications in Belgium, 1946-2007

The remainder of this paper is structured as follows. In section 2, the field of information retrieval (IR) is explained, with a focus on the retrieval of legislation. In section 3, one specific IR technique, classification schemes, in which persons can choose a certain class of legislation and get an overview of relevant laws and articles for that class, is discussed. In section 4, we discuss different ways of creating and maintaining the mapping between laws and class labels, and conclude that legislators should add class labels to the legislation they create. A short summary, conclusions and directions for future research are written in section 5.

2. Information Retrieval

Information Retrieval (IR) is a subfield of computer science, concerned with finding information in a a large data collection. More formally, it is "the process that selects documents relevant to a user's information needs out of a (large) document collection" [4, 5]. A typical basic IR system allows the user to enter a *search query*, composed of *keywords* and *modifiers* such as AND and OR [4]. The system then searches documents that contain the desired words. Over time, the capabilities of this kind of IR systems have improved greatly. Nowadays, such systems can handle user queries in a more reliable way, and are often able to recognize and deal with issues as synonyms, word inflections, and typing errors [4]. Search results are often ranked according to some relevance measure, and sometimes even automatically clustered in meaningful result groups [5].

Despite these improvements, this kind of IR systems suffers from a fundamental problem: they expect the user to come up with good keywords. While this is not a big problem when a user searches for a easily defined general topics as "football" or "Barcelona", it is for complex legislative concepts such as "Enforcement", since these complex legal concepts do not correspond with explicit wording in document texts [6]. Also, while in general IR tasks the goal is to find some or most relevant information, in a legal setting, one should be able to find all, or almost all legislation relevant to a certain issue [6]. If a legislative IR tool failed to retrieve only one crucial article for a certain query, the tool would actually do more harm than good for that query. Therefore, it becomes clear that keyword search is not very appropriate in the context of the retrieval of legislation. Thus, more advanced IR schemes are necessary for the retrieval of legislation [6, 7]. In our opinion, the only viable solution to these problems is the use of a classification scheme, in which laws and articles are classified into legal topics. That way, users can browse through these predefined classes of law, and can thus easily find all legislation related to their specific issue. This kind of classification is discussed in more detail in the following section.

3. Classification schemes to access legislation

The use of a classification scheme boils down to tagging articles with class (index) labels, and consequently offering users the possibility of selecting labels and getting a list of related articles as a result.

Good classification schemes have three structural properties. First of all, indexes can be organized into a hierarchic structure. For example, the main index "Enforcement" could be subdivided into "Criminal enforcement", "Civil enforcement", and "Administrative enforcement". Second, the classification can happen along more than one axis. For example, there could be a thematic axis, dividing law into topics such as "Tax Legislation", "Environmental Legislation", and "Energy Legislation", and a legal axis, with topics such as "Enforcement" or "Procedures". Every article then is assigned a label for each axis. A third, and final property is that articles can also have different index labels within one axis. For example, an article could contain information about both "Civil enforcement" and "Administrative enforcement", and should therefore receive both labels.

For users, such a classification scheme offers great usability and access. Even without knowing the specific terminology of a certain legal domain, one can easily browse through the different levels of the classification hierarchies to find the legislation of his or her interest.

This plea for a classification approach to legislative information retrieval is not new (e.g. [6, 8]). Nevertheless, despite the advantages, this classification approach is still not widely adopted. Most of the currently existing initiatives offering online access to legislation still only offer simple keyword search, often even without the improvements discussed above, aggravating the problems of this kind of search. This slow adaptation of classification schemes is due to one big disadvantage of such scheme, namely that the creator of the classification scheme has to classify articles into the classification scheme. In existing systems, this classification task is always the responsibility of the IR system builders and content managers. This approach is, however, as we will argue hereafter, a flawed one.

Usually, these content managers employ legal experts to classify the articles manually [6]. This is prohibitively slow and expensive, especially because of the high rate at which new legislation is created. On top of that, even legal experts make mistakes. An obvious alternative is to automate this task. This approach, while cheap and fast, has problems of its own. While text classification is a mature domain, with very good results in e.g. the automated classification of news articles in classes such as "Politics" or "Sport" [9], achieving accuracy rates of 90%+, the approach does not work very well for legal topics [6, 10]. This is mainly because legal concepts are too abstract to be easily learned by a classifier. In the past, there was always optimism that this problem would be overcome, but now we have indications that this will not be the case in the near feature [11].

Thus, while the classification approach is necessary from a user perspective, it is very hard to create such a classification if the content managers have to add the class labels *after* the legislation is created. Therefore, the only feasible option is that legislators themselves tag new legislation with appropriate labels, a path we pursue in the next section.

4. Legislators should do the classification

As there is no practical way to classify legislation after it is issued, legislation should be tagged with class labels at the time it is created. The people most suited to do this would be the legislators themselves. As they are obviously very familiar with the legislation they create, the extra effort required to tag it with appropriate thematic and legal class labels should be negligible.

Of course, the tagging of new legislation does not solve the problem of legacy legislation. However, as shown in Figure 1, the amount of new legislation is ever-increasing, and thus poses the greatest challenge in the long term. Also, indexing all existing legislation is a huge task but, at least, a one-time and finite effort. This is strongly opposed to the continuous, never-ending task of indexing all new legislation.

5. Conclusions

In this paper, we argue that legislators should tag every piece of legislation they produce with class labels. While the idea of adding class labels to legislation is not new, here we have shown that doing it *after* legislation is issued is practically impossible, whether done manually or automatically. Also, we have argued that adding class labels to legislation is not only desirable, but absolute necessary to safeguard the ability of persons to fulfill the adage *Nemo censetur ignorare legem.*

References

[1] Laurence D. Houlgate. Ignorantia juris: A plea for justice. *Ethics*, 78(1):32–42, 1967.

[2] John Austin. *Lectures on Jurisprudence*, pages 496–501. London, 1873.

[3] H. Matthijs. To appear.

[4] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval.* Addison Wesley, 1st edition, May 1999.

[5] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval.* Cambridge University Press, 1 edition, July 2008.

[6] Marie-Francine Moens. Innovative techniques for legal text retrieval. *Artif. Intell. Law*, 9(1):29–57, 2001.

[7] J. P. Dick. Conceptual retrieval and case law. In *ICAIL '87: Proceedings of the 1st international conference on Artificial intelligence and law*, pages 106–115, New York, NY, USA, 1987. ACM.

[8] C. Biagioli. Towards a legal rules functional micro-ontology. In *Proc. of workshop LEGONT '97*, 1997.

[9] F. Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002.

[10] R. Opsomer, G. De Meyer, C. Cornelis, and G. Van Eetvelde. Exploiting properties of legislative texts to improve classification accuracy. *To be published in proceedings of Jurix 2009*, 2009.

[11] R. Opsomer, G. De Meyer, and G. Van Eetvelde. Automated classification of legal texts will never work. *TO BE WRITTEN AND PUBLISHED*.