The OpenLaws project: Big Open Legal Data

Winkels, R.

[Link to publication](#)

THE OPENLAWS PROJECT:
BIG OPEN LEGAL DATA

Radboud Winkels[1]

[1]Associate Professor, University of Amsterdam, Leibniz Center for Law.
PO Box 1030, 1000 BA Amsterdam, NL
winkels@uva.nl; http://www.leibnizcenter.org/~winkels/

Keywords: *Network analysis, recommender system, open data.*

Abstract: *In the OpenLaws project we aim to deliver a platform that enables users to find legal information more easily, organize it the way they want and share it with others. Together with users we intend to create a network of legislation, case law, legal literature and legal experts – both on a national and a European level – leading to better access to legal information. Besides involving users, we also look at semi-automatic ways of connecting sources of law through the use of natural language techniques and network analysis. In this contribution we will present first results for combining open sources of law in the Netherlands. Not only can we present related legislation and case law, but the law citations in case law influence the ranking of relevant legislation*

1. Introduction

More and more sources of law are freely available online in Europe and the rest of the world. Most of the time however, these are stand-alone web services or databases, containing one type of documents, not linked to other sources. For instance the Dutch portal for case law – rechtspraak.nl – contains a (small) part of all judicial decisions in the Netherlands. Case citations in these decisions are sometimes explicitly linked, references to legislation are not. From earlier research we know that professional users of legal documents would like to see and have easy access to related ones from other collections. E.g. when we evaluated a prototype system that recommends other relevant articles and laws to users of the official Dutch legislative portal, they told us they would like to see relevant case law and parliamentary information as well (Winkels e.a. 2013).

This is one of the aims of the OpenLaws project, delivering a platform that enables users to find legal information more easily, organize it the way they want and share it with others (Wass e.a., 2013). Together with users we intend to create a network of legislation, case law, legal literature and legal experts – both on a national and a European level – leading to better access to legal information. The OpenLaws Internet platform is based on open data, open innovation and open source software. This means that OpenLaws integrates legal communities and IT communities in the design and in the operation of the system.

Besides involving users, we also look at semi-automatic ways of connecting sources of law through the use of natural language techniques and network analysis. In this contribution we will present first results for combining open sources of law in the Netherlands. Not only can we present related legislation and case law, but the law citations in case law influence the ranking of relevant legislation.

2. Big Open Legal Data (BOLD)

Legal data, narrowly conceived, consists of two types of BOLD objects:

1. structured texts, hierarchically decomposable into chains of text fragments, and

2. metadata about texts and text fragments;

    a. labeled links between texts fragments, and

    b. arbitrarily complex features of texts and text fragments.

Bibliographic convention (Sauer, 1998) is to distinguish legal texts and text fragments on at least four levels:

1. On the *item* level legal texts and text fragments can be dereferenced by identifier and copied, resulting in a new item;

2. On the *manifestation* level any change to the data produces a new manifestation, including a change of data format, annotation of structure, or the embedding of metadata;

3. On the *expression* level only a change of the text by its author produces a new expression;

4. On the *work* level a text is identified by the details of its publication: as long as the title, author, and publication date remain the same, expressions are versions of the same work.

Enrichment of legal data consists of 1) creating alternative manifestations of an expression, or 2) adding metadata about an expression or work. One work may be expressed multiple times. An expression may have many manifestations. A manifestation may be copied to many items.

2.1. Versioning

For regulatory text, the distinction between works and expressions is of critical importance, because the text is typically changed over time. Most works are decomposable into a single versioning chain of expressions. In some cases (retroactive annulment of changes and unforeseen changes to scheduled changes in the future) the versioning chain may change over time retroactively, resulting in alternative chains of versions of a text.

2.2. Languages

Many works are moreover available in alternative language variants. Support of this is obviously an important requirement in the EU.

2.3. References

Conventionally, regulatory text refers to other regulatory text on the work level: which version should be used is left to the reader (see Figure 1). Some call this a *dynamic* reference. A court decision refers to a specific version of a regulatory text on the expression level. Some call this a *static* reference. Any other text that refers to legislation by default refers to a specific version, unless the text is under editorial control and guaranteed to be up to date with the text it refers to. Obviously, texts may discuss an old version, compare an old version to a new version, compare two language variants of a version, or (very frequently) discuss an anticipated version of a regulatory text.
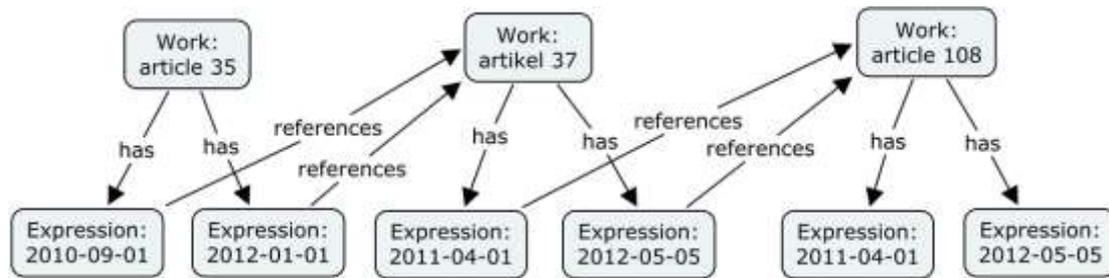
Figure 1: Example of dynamic references in regulatory text. The work ´article 35´ has two expressions; both refer to the work level of article 37. The actual references can only occur in a manifestation of an expression (e.g. an XML file) and be seen by a user in a physical item on for instance her computer.

2.4. Mixed content and quoting

Legal text is often quoted, in modifying legislation, in court decisions, papers and books, etc. Quoting is an alternative way of referencing information, and the quoted text fragment is both a part of the quoted and of the quoting document. It is moreover a potential source of interesting and innovative manifestations of text fragments.

3. Dutch Official Portals for Law

There are several official portals for legal data in the Netherlands. The two most important ones are the one for laws and treaties (wetten.nl) and the one for case law (rechtspraak.nl).

3.1. Wetten.nl

The official Dutch portal for national legislation allows users to search and browse all legislation as text with hyperlinks. When an article is in focus on that site, the structure (e.g. chapters and paragraphs) of the regulation is shown on the left hand side, but it does not clearly show the chapter you are in (Figure 2). It is also not easy to switch to earlier versions of the same article or find out which other sources of law refer to the particular article in focus.
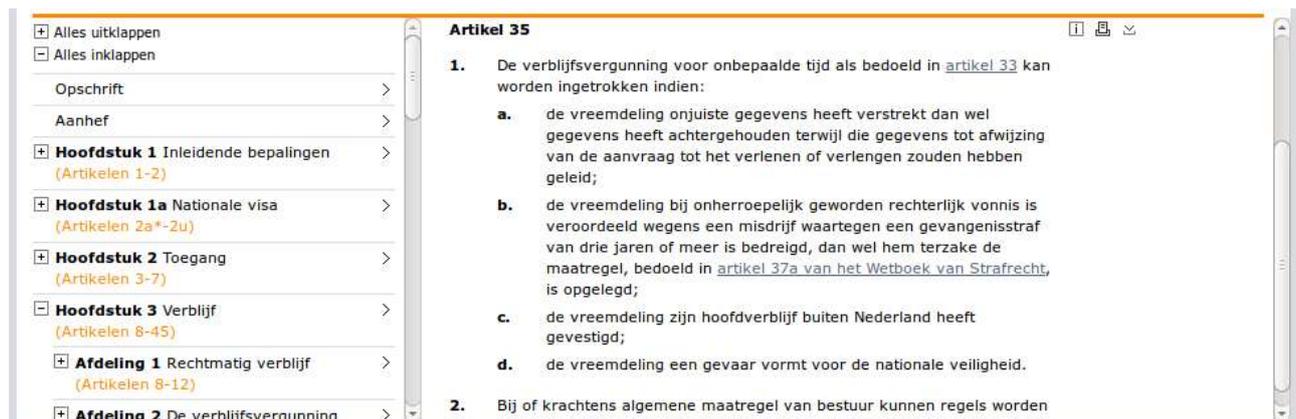

Figure 2: Interface of Dutch legislative portal. The user is focussing on article 35 of immigration law.

3.2. Rechtspraak.nl

The Dutch portal for case law contains a (small) part of all judicial decisions in the Netherlands. Case citations in these decisions are sometimes explicitly linked, references to legislation only the

main one(s) in recent cases (see Figure 3).[1] The texts are available in an XML format, basically divided in paragraphs using `<para>` tags, with a few metadata elements. The most relevant metadata for our purpose are:

- – The date of the decision ('Datum uitspraak')
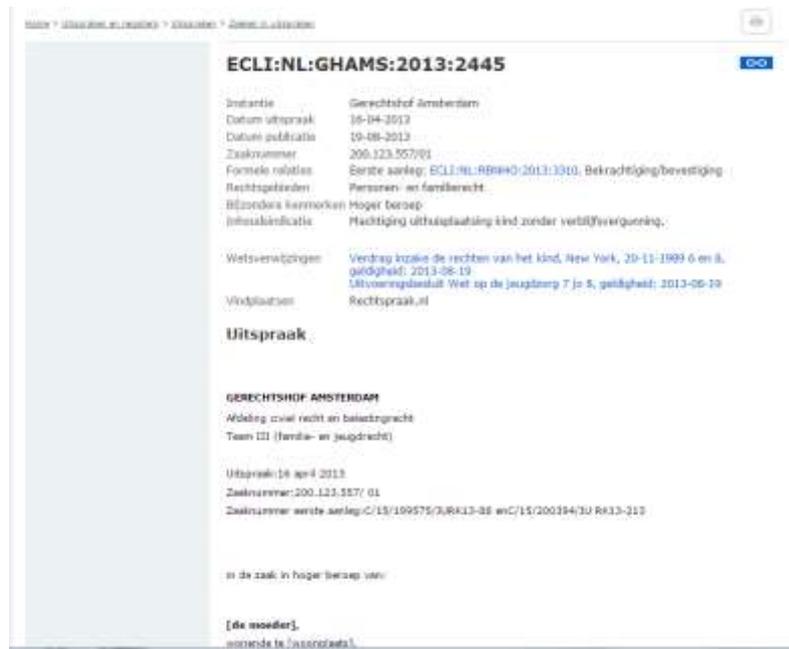- – The field(s) of law ('Rechtsgebieden')
- – The court ('Instantie').

Figure 3: Interface of case law portal.

4. Creating a Network of Legal Data

We want to create a network of sources of law that contains both legislation and case law. Later on we would like to add legal doctrine and commentaries as well. In OpenLaws.eu we envision this in two ways: We will enable users to connect and organize sources of law and we will try to do this automatically.

4.1. User initiated linking

The OpenLaws.eu project will deal with the following challenges in managing user communities:

- – Permitting user communities to freely organize themselves, to freely share information with the agents of their choice, excluding others, and to freely exclude information from view that they deem irrelevant or of low quality;
- – On the other hand, providing incentives to share information that they produce as widely as possible, as implied by the open legal data concept;
- – And finally, determining the relevance of information for specific agents, addressing both the general quality of information (or the confidence in the skills of its producer) and specific agent information needs.

---

[1] In a metadata element in the header of the document that contains the 'main' article(s) for the decision, not in the running text.

While the second challenge above is arguably the key objective of the project, the first challenge is important to the business case for the framework and of central importance to addressing the third challenge, safeguarding the quality of BOLD in the long run.

4.1.1. Folders, shopping carts, and shopping lists

Users looking for legal information require some way to select and keep it. One may think of this in terms of the shopping cart metaphor. By storing the contents of a shopping cart in a folder the user enriches the text fragments in the shopping cart with features. The folder is a set based on shared features.

Alternatively, the user may use the shopping cart as a basis for producing a new text, quoting the selected text fragments. In a new text the fragments are ordered in a chain; the shopping list, whose order may be changed, is a more appropriate metaphor for this use case.

The text fragment is the fundamental unit of the user interface: (named) containers, like article 1, paragraph-sized blocks, or sentences.

An Interface challenge will be to clearly distinguish the various bibliographic levels of the text fragments as indicated above. It should e.g. be clear to the user and to the system whether the user is putting a 'work', or a specific version ('expressions') in the shopping cart.

4.2. Automatic Linking of Sources of Law

The data available in the Dutch official portals is not ideal for further analysis, e.g. natural language parsing or network analysis. For legislation we started transferring the data to an RDF linked data format in 2011 and store it at the Metalex Document Server[2] (Hoekstra, 2011). The server currently contains 41,581 document versions (January 2015), and this number is growing every day since every change to the wetten.nl site is added to the triplestore. In Winkels e.a. (2013) we describe how we generate networks for legislation using SPARQL queries on the RDF data.

As stated above, the court decisions published at the official portal do not contain (inline) explicit, machine readable links to cited legislation. Neither do they contain machine readable links to *all* cited case law.[3] We resort to parsing techniques to make these citations explicit and count them. In Winkels e.a. (2011) we describe how we built the network of Dutch case law.

Here we will describe how we find and resolve the references to legislation in case law and how we create an integrated network of Dutch sources of law and use this to suggest possibly interesting documents to users of a legislative portal. We chose to work with a subset of all case law to start with; those cases that were tagged as belonging to `immigration law`. That gave us 13,311 documents to work with.

4.2.1. Locating and resolving references

For locating references to legislation in case law we use regular expressions as we have done in the past together with a list of names and abbreviations of Dutch laws (de Maat e.a., 2006). This list also contains the official identifier of the law (the BWB-number), which can be used for resolving the reference later on. We consider high precision to be more important than high recall. Users will forgive us if we miss a reference, but be annoyed by false ones.

We evaluated this procedure by checking 25 randomly selected documents by hand. These documents contained 163 references to legislation of which 141 were correctly identified (recall of

---

[2] http://doc.metalex.eu/
[3] The more recent cases contain more explicit links.

87%). There was one false positive (precision of 99%). The references we missed were mostly those to the 'Vreemdelingencirculaire' (a lower law that has a different structure than regular ones) and treaties with very long names like 'Europees Verdrag tot bescherming van de rechten van de mens en de fundamentele vrijheden' . If we tried to capture this with the regular expression, the regular expression would match too easily, often matching entire sentences where it should have matched only the law. We declared these conventions outside the scope of this first experiment.

Resolving the references was a bit trickier, since sometimes they used anaphora, e.g. referring to 'that law'. In that case, the citation was resolved by using the previous law identifier if it existed. We used the same process for resolving ambiguous title abbreviations; e.g. `WAV` is an abbreviation of 'Wet Arbeid Vreemdelingen', 'Wet Ammoniak en Veehouderij' and 'Wet Ambulancevervoer'. Most of the time the full title is used before the abbreviation is used. Another issue is determining the exact version of the law the case refers to. Typically, a judge will refer to the version that is in force at the moment of the decision, but it may also be the version that was in force at the time of the relevant facts, or even sometimes an earlier version of the relevant law, etc. We cannot decide which version is the correct one without interpreting the content of the case. Therefore we decided to resolve the reference to the work level of the source of law, i.e. no particular version. The resulting references are added to the XML of the case law document. The final network of the 13,311 case documents has 85,639 links to legislation (on average 6.5 references per case); the links connect the ECLI identifier[4] of the case with the BWB identifier of the source of law.

We evaluated the resolving process by checking 250 random ones of all the references found, by hand. Of these, 234 should have been resolved since the other 16 were outside of the scope of this experiment. 198 were resolved correctly (a recall of 85%). We had 10 'false' positives, i.e. references that were declared out of scope, so a precision of 95%. The results were good enough to continue.

Since case decisions may refer to the same source of law, e.g. an article, more than once, we count the number of references and compute the weight of the link between the case and the article as: $W = \frac{1}{n}$ where n is the amount of occurrences of a certain reference and W is the weight of the edge. The lower the weight, the stronger the impact on the network is.

5. A Prototype Legal Recommender System for Dutch Law

When a user clicks an article in the legislative portal, related case law and legislation is retrieved:

1. The system checks whether the article appears in the case law network. If so, it creates a so-called ego graph, a local network containing all the nodes and edges within a certain weighted distance from the current node (Newman, 2010). We start searching with a weighted distance of 0.4 and gradually increase it up to 2.0 until we have a sufficiently large, but still manageable network.

2. To find relevant legislation, the system also checks whether the current node is in the legislative network of the MetaLex document server. If so, it again creates an ego graph, this time for an unweighted network. To control the size of the graph, we use only references coming from the selected version (expression) of the current node.

3. If we have two local networks, we want to combine them in order to (better) predict the importance of legislative nodes. To do this, we need to assign weights to the legislative

---

[4] 'European Case Law Identifier'; see Council conclusions on ECLI at: http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52011XG0429(01)

graph. We chose the value 0.1 as it allows the legislative network to influence the result but not overrule the case law references.

4.   Finally, we use betweenness centrality on the combined network to determine the most relevant articles for the current focus. The betweenness centrality of a node is the sum of the fraction of all-pairs shortest paths that pass through that node. The results are shown to the user in the top of frame A of Figure 4.
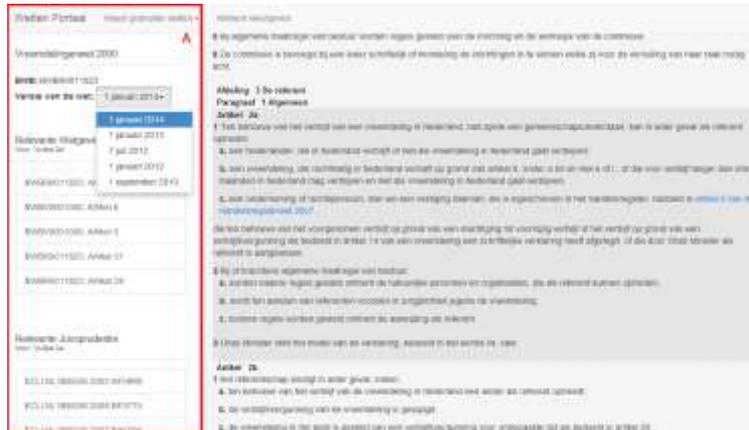
Figure 4: The prototype legal recommender system. The user has article 2a of the Immigration law in focus. Current version is January 2014 (see pull-down menu at left). Relevant other articles are presented in window A (red or dark border) on the left and below that relevant case law.

5.1.   A First Evaluation

We asked several professional users of the Dutch Immigration Service to use the prototype system and fill in an evaluation form afterwards. Due to the holiday season, only 3 users replied so far, but they were positive. They appreciated the clean and uncluttered interface, indicated that it was easy to understand without help and liked the indication of the number of times a case was referring to an article when you click on a case. They also noted that the inclusion of case law added suggestions of relevant articles that were not available in a previous system (Winkels e.a., 2013). They complained about the slowness of the system and the fact that references to the 'Vreemdelingencirculaire' were missing (as we explained above).

6. Conclusions

We described first results of research that is part of the OpenLaws.eu project. Ultimate aim is to deliver a platform for using, sharing and enriching big open legal data. Besides offering users the opportunity to collect, organize and annotate legal data, we also want to offer automatic suggestions based on analysis of existing data. In the future that may also be based on user generated data, but for now it is based on the analysis of official sources of law. We have presented first results for available data in the Netherlands. Based on analysis of the network of both legislation and case law, we offer users suggestions for interesting material based on their current focus document.

We have shown that it works quite well to automatically find and resolve references to legislation in Dutch case law. These references can be used to provide users of the legislative portal with relevant judicial decisions given their current focus and moreover, suggest additional relevant legislative sources. The parser can easily be improved a little and the prototype system will perform much better when running on a proper server. We also did not exploit the network of case law itself this time. In Winkels e.a. (2011) we showed that this can be used to estimate the authority of cases, so if we include this the suggestions of relevant case law should be improved.

Another next step is adding legal commentaries and doctrine to the network and possibly parliamentary data.

7. Acknowledgements

8. References

Hoekstra, R., The MetaLex Document Server - Legal Documents as Versioned Linked Data, pp. 128–143. Springer (2011).

Maat, E. de, Winkels, R., and Engers, T. van., Automated detection of reference structures in law. In T. van Engers (ed), JURIX 2006, IOS Press, Amsterdam, pp. 41-50 (2006).

Newman, M., Networks: An Introduction. Oxford, England: Oxford University Press (2010).

Saur, K.G., Functional requirements for bibliographic records. UBCIM Publications - IFLA Section on Cataloguing, 19 (1998).

Wass, C., Dini, P., Eiser, T., Heistracher, Th., Lampoltshammer, Th., Marcon, G., Sageder, C., Tsiavos, P. and Winkels, R., OpenLaws.eu. Proceedings of the 16th International Legal Informatics Symposium IRIS 2013, Salzburg, Austria (2013).

Winkels, R.G.F., Boer, A. and Plantevin, I., Creating Context Networks in Dutch Legislation. In K. Ashley (ed). Legal Knowledge and Information Systems. JURIX 2013. IOS Press, Amsterdam, pp. 155-164 (2013).

Winkels, R.G.F., de Ruyter, J. & Kroese, H., Determining Authority of Dutch Case Law. In K. Atkinson (ed). JURIX 2011: The Twenty-Fourth International Conference. Volume 235 of Frontiers in Artificial Intelligence and Applications, IOS Press, Amsterdam, pp. 103-112 (2011).