WORKSHOP
“NETWORK ANALYSIS IN LAW”

FRIDAY JUNE 14TH 2013
IN CONJUNCTION WITH ICAIL 2013:
XIV INTERNATIONAL CONFERENCE ON AI AND LAW

RADBOUD WINKELS (EDITOR)
Preface

This workshop aims to bring together researchers from computational social science, computational legal theory, network science and related disciplines in order to discuss the use and usefulness of network analysis in the legal domain.

Two obvious strands of research come to mind:

1. **Analysing and visualizing networks of people and institutions**: Law is made by people, about and for people and institutions. These people (or institutions) form networks, be it academic scholars or criminals and these networks can be detected, mapped, analysed and visualised;

2. **Analysing and visualizing the network of law**: Law itself forms a network. Sources of law refer to other sources of law and together constitute (part of) the core of the legal system. In the same way as above, we can represent, analyse and visualise this network.

A third area of research is where these two networks meet:

3. **People or institutions create sources of law or appear in them**: Research on the network of one may shed light on the other. Two examples:
   a. Legal scholars write commentaries on proposed legislation or court decisions. Sometimes they write these together. These commentaries may provide information on the network of scholars; the position of an author in the network of scholars may provide information on the authority of the comment.
   b. ‘Criminals’ appear in court decisions and may appear in more than one. Information on the network of criminals may help in finding related cases and decisions. Criminals that appear together in a court case may help in building the network of criminals.

We invited papers on and demonstrations of original work on these and other aspects of network analysis in the legal field. Submissions were subject to a light review process on appropriateness for this call, originality of the research described and technical quality.

Workshop Format
Short, 15 minutes presentations and/or demonstrations followed by discussion.

Organizing Committee
Michael Bommarito, ReInventLaw Laboratory, Michigan, USA
Romain Boulet, University of Lyon, France
Daniel Katz, Michigan State University, USA
Marc Lauritsen, Capstone Practice, USA
Nicola Lettieri, University of Sannio Law School, Italy (co-chair)
Bill Speros, Speros & Associates LLC, USA
Innar Liiv, Tallinn University of Technology, Estonia
Thomas Smith, University of San Diego Law School, USA
Radboud Winkels, Leibniz Center for Law, Netherlands (chair)
## WORKSHOP PROGRAM

Every presentation should be 15 minutes max, followed by 10 minutes for questions and discussion.

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>09.00 – 09.05</td>
<td>Opening&lt;br&gt;Radboud Winkels and Nicola Lettieri</td>
</tr>
<tr>
<td>09.05 – 09.30</td>
<td>Bad law before it goes bad: Citation networks and the life cycle of overruled Supreme Court precedent&lt;br&gt;Ryan Whalen</td>
</tr>
<tr>
<td>09.30 – 09.55</td>
<td>Knowledge Network Based on Legal Issues&lt;br&gt;Paul Zhang, Harry Silver, Mark Wasson, David Steiner and Sanjay Sharma</td>
</tr>
<tr>
<td>09.55 – 10.20</td>
<td>Network Analysis of Manually-Encoded State Laws and Prospects for Automation&lt;br&gt;Patricia Sweeney, Elizabeth Bjerke, Margaret Potter, Hasan Guclu, Keane Christopher, Kevin Ashley, Matthias Grabmair and Rebecca Hwa</td>
</tr>
<tr>
<td>10.20 – 10.45</td>
<td>From structure to function: exploring the use of text and social network analysis in criminal investigations&lt;br&gt;Nicola Lettieri, Delfina Malandrino, Raffaele Spinelli and Carlo Rinaldi</td>
</tr>
<tr>
<td>10.45 – 11.10</td>
<td>A Network Analysis of Dutch Regulations - Using the MetaLex Document Server&lt;br&gt;Rinke Hoekstra</td>
</tr>
<tr>
<td>11.10 – 11.30</td>
<td>Coffee Break</td>
</tr>
<tr>
<td>11.30 – 11.55</td>
<td>A Text Analysis Framework for Automatic Semantic Knowledge Representation of Legal Documents&lt;br&gt;Ákos Szőke, Krisztián Mácsár and György Strausz</td>
</tr>
<tr>
<td>11.55 – 12.20</td>
<td>Graph-Based Linking and Visualization for Legislation Documents (GLVD)&lt;br&gt;Dincer Gultemen and Tom van Engers</td>
</tr>
<tr>
<td>12.20 – 12.55</td>
<td>Finding and Visualizing Context in Dutch Legislation&lt;br&gt;Radboud Winkels and Alexander Boer</td>
</tr>
<tr>
<td>12.55 – 13.00</td>
<td>Closing Remarks</td>
</tr>
</tbody>
</table>
# Table of Contents

Bad law before it goes bad: Citation networks and the life cycle of overruled Supreme Court precedent .................................................. 1  
  *Ryan Whalen*

Knowledge Network Based on Legal Issues ........................................ 2  
  *Paul Zhang, Harry Silver, Mark Wasson, David Steiner and Sanjay Sharma*

Network Analysis of Manually-Encoded State Laws and Prospects for Automation............ 3  
  *Patricia Sweeney, Elizabeth Bjerke, Margaret Potter, Hasan Guclu, Keane Christopher, Kevin Ashley, Matthias Grabmair and Rebecca Hwa*

From structure to function: exploring the use of text and social network analysis in criminal investigations..................................................... 4  
  *Nicola Lettieri, Delfina Malandrino, Raffaele Spinelli and Carlo Rinaldi*

A Network Analysis of Dutch Regulations - Using the MetaLex Document Server........... 5  
  *Rinke Hoekstra*

A Text Analysis Framework for Automatic Semantic Knowledge Representation of Legal Documents........................................................................ 6  
  *Ákos Szőke, Krisztián Mácsár and György Strausz*

Graph-Based Linking and Visualization for Legislation Documents (GLVD)..................... 7  
  *Dincer Gultemen and Tom van Engers*

Finding and Visualizing Context in Dutch Legislation.......................................... 8  
  *Radboud Winkels and Alexander Boer*
Bad law before it goes bad: Citation networks and the life cycle of overruled Supreme Court precedent

Prepared for the ICAIL 2013

Abstract

This essay expands the empirical analysis of overruled Supreme Court decisions by focusing on citation network centrality measures and comparative depreciation analysis to compare overruled decisions to matched peers. We demonstrate that those decisions which go on to be overruled tend to occupy more central citation network positions, cite more central cases, and depreciate at a slower rate than similarly influential decisions. These empirical demonstrations of how bad law is distinct in both the way it cites and in the way it is cited help shed light on how the judiciary affects legal change. We see that judges reserve the power of overrulings for decisions that are both significantly more central than their peers and for decisions that remain salient for longer than their peers. This not only sheds light on our understanding of overruled precedent, but also contributes new network-analytic approaches to legal research.
By definition, every precedent now deemed bad law was at some point in time good law. While some cases are overturned almost immediately after they are decided and thus have little opportunity to exert any influence on subsequent decisions, the majority remain good law for years during which time they are eligible to be cited as authoritative sources of law. We can think of the period between when a decision that will go on to be overturned is originally issued and when it is overturned as its lifespan. During this period, these precedents are eligible sources of law, and each will have a unique biography that tells the story of how it was used – or not used - by the legal system before eventually being overturned.

Prior studies of overturned precedent have almost exclusively focused on the two most obvious instances of judicial interaction with good/bad law: the writing of the original decision and subsequently the decision that ultimately overturns it.\(^1\) However, good/bad laws often exist for decades before they are overturned (Brenner & Spaeth, 1995), during their tenure as good law they help shape the law that itself helps shape society. Situating precedent within the entire network of legal decisions and analyzing its centrality and influence over time can help us better understand the differences between good and bad law. To do so, we must first identify which traits and network centrality attributes to measure.

**Depreciation:** One aspect of precedent that seems relatively universal is the fact that as a decision ages its influence within the legal system depreciates (Landes & Posner, 1976). Black and Spriggs demonstrate that, regardless of some initial differences in the rate at which precedents depreciate, after about a decade almost all cases attract citations at a similarly decreasing rate (Black & Spriggs II, 2009). However, no study has compared the depreciation rates of cases that go on to be overturned to those that do not.

---

\(^1\) See e.g. (Ulmer, 1959) (Schmidhauser, 1961) (Brenner & Spaeth, 1995), (Banks, 1991)
There are three possible ways overruled cases could depreciate vis-à-vis cases that remain good law. One, they could depreciate at the same rate. This would suggest that there is no quantitative difference in the ways good and bad laws are cited. Two, overruled cases could depreciate more quickly than non-overruled cases. This could come about if there is generally a delay between a case being recognized as bad law and its eventual overruling. In the interim, judges may be less inclined to cite the bad law, thus leading to faster depreciation. Three, overruled cases could depreciate more slowly than their non-overruled counterparts. This outcome could come about if the issues covered by overruled laws increase in salience as an overruling becomes imminent. Increased salience or controversy around cases that will be overruled could lead to citations that distinguish the bad law, thus putting off an overruling. In order to understand better which of these three scenarios is in fact the case, we ask research question one:

**RQ1: Do Supreme Court decisions that go on to be overturned have different rates of depreciation than decisions that do not?**

**Centrality.** Fowler and Jeon demonstrate that both inward and outward network centrality are excellent predictors of whether or not experts deem a case important (2008, p. 23). They also show that the inward authority scores of overruled cases tend to be much higher than the global average (p. 25). However, this comparison is somewhat unfair as it includes in the global average all those thousands of cases that receive zero or close to zero citations. A better comparison would match overruled cases against cases that receive a similar number of citations. This would allow us to know more definitively how the importance of overruled decisions compares to that of cases with a similar profile.

If both overruled and non-overruled cases have similar authoritativenss, it would suggest that authority or importance is not a distinguishing feature of bad law. If overruled cases
have higher authority scores, it would suggest that bad laws tend to come from important or “hot” legal areas. Finally, if the opposite is true and overruled precedents have lower authority scores than similar counterparts it would suggest either that bad laws are seen as such and thus less likely to attain those very high centrality scores, or alternately that judges are disinclined to overrule cases in particularly central positions – perhaps because their authority lends them such gravitas that overruling them is much more difficult. To better understand which of these scenarios is the case we pose research question two:

\textit{RQ2: Are Supreme Court decisions that go on to be overturned cited by more or less authoritative law than similar counterparts?}

The flip side to how authoritative the laws that cite a case are is how authoritative their own body of cited precedent is. Fowler and Jeon define a “hub” score that takes into account the network position of each precedent cited to determine how much outward authority each decision has. We can examine these scores to determine whether or not those laws that go on to be overruled cite precedents that are themselves more or less authoritative than those cited by similar cases. As with the authority scores, using a global average is not the best comparison measure. Rather, because hub authority is proportional not only to the authority of cases cited but also to the number of cases cited, we must match overruled cases with decisions issued in the same year that cite a similar number of precedents and then compare the overruled cases to their matches.

If we find that the hub authority of overruled cases is not significantly different from that of their matches, we can infer that the strength of the precedential justification for a decision has little relationship to whether it is good or bad law. If instead we find that overruled cases have greater hub authority than their matches it could perhaps suggest that contentious laws feel the need to buttress their position by citing widely and authoritatively. Finally, if cases
that go on to be overruled have lower hub authority than their matches, it could be that the relative weakness in precedential justification for these findings contributes to their eventual overturning. This leads to research question three:

\textit{RQ3: Do Supreme Court decisions that go on to be overturned rely on more or less authoritative law than their counterparts?}

\textbf{Method.} The Supreme Court citation and precedent centrality data used in the analysis below comes from a set provided by Lexis-Nexis and used originally in Fowler et al’s (2007) analysis of precedent centrality measures. It includes complete data on citations between Supreme Court cases from 1791 to 2005. The Government Printing Office report on overruled decisions was used to identify Supreme Court decisions that go on to be overturned and the decisions that overturn them ("Supreme Court Decisions Overruled by Subsequent Decisions," 2002).

Because this study aims to expand the analysis of overruled precedent beyond the point of original decision and overruling decision, we choose to focus on cases that survive at least ten years before being overruled. Precedents that are overruled quickly have fewer opportunities to influence the legal system. While important in their own manner, these flash-in-the-pan precedents are qualitatively different from precedents that exist for decades as good law before they are overruled.

\textit{Depreciation and authority matches.} Each overruled case is matched to two other decisions written in the same year. In order to find suitable matches for depreciation and authority score comparisons we calculate the total number of citations received by each overruled decision before it is overruled. We then choose as a match that decision written in the same year that has the smallest absolute value difference in citations received over the same period.
Matching cases by the number of citations received provides a set of matches that we can expect to be reasonably similar in importance to the set of overruled cases. There will of course be differences between the area of law covered by the overruled case and its match, but there is little reason to expect that these differences are systematic. Given a large enough sample of overruled cases and their matches, the primary distinction between them will be that the overruled cases go on to become bad law, whereas the matches do not. Thus, any significant differences we observe between the two populations are most likely to be due to this underlying difference.

*Hub matches.* Because hub authority is a function of both the authority of cited cases as well as the number of cases cited, matching based on in citations does not provide an adequate comparison. In order to compare hub scores of overruled cases with those of cases that do not go on to be overruled, we select a second collection of matched cases. In this second round of matching, each overruled case is matched the another decision written in the same year that has the lowest absolute value difference in out degree.

*Hub/Authority scores.* Fowler *et al* use Kleinberg’s (1999) hub/authority measures to calculate scores for each case. These scores rely upon the number of citations received by each case as a measure of their importance. A hub score is a measure of *inward* importance. It is calculated by summing the authority scores of each case that cites the case in question. An authority score is a measure of *outward* importance. It is calculated by summing the hub scores of each case cited by a given case. Hub and authority percentile scores are used below as we track the importance of overruled cases across time.

As new cases are decided, the citations they produce change the network structure, in turn changing both hub and authority scores. We can monitor these changes over time to create hub and authority curves that track the importance of the cases in our study across time. The
data used below tracks yearly hub/authority scores for both overruled and matched cases. These scores are then used to create mean hub and authority curves for each type of case.

**Depreciation:** To measure depreciation we sum the citations each case receives on a yearly basis. This process creates a biography of each case, with entries for each year of age that tell us how often that case was cited when it was x years old. We can then use these biographies to create mean depreciation curves for both overruled cases and their matches.

**Rescaled depreciation.** Comparing raw depreciation curves is problematic for two reasons: there is great variation in the number of citations cases receive and there is also great variation in the length of time overruled cases survive before they become bad law. We deal with these challenges by rescaling both the x and y axis of each case’s biography.

The x axis represents age. In order to compare cases that live for different lengths of time, each case’s lifespan is rescaled from 0-10. A citation received by a case immediately after it is issued counts toward the 0 point, while decisions immediately prior to the overruling decision count towards the 10 point and so on. The overruling decision is not included in the depreciation curves.

The y axis of our case biographies represents the number of citations received. To deal with scale variance of this measure we rescale the y axis to instead represent the proportion of total citations received before the case in question is overruled – or in the case of matches before its match is overruled. So, if a case receives 30% of the total citations it will receive before being overruled immediately after it is written, it will score (0.3,0) in (x,y) terms.

**Significance tests.** The data assembled above provide three ways to compare overruled cases with their matches: hub score, authority score, and depreciation. In order to determine whether or not there are statistically significant differences between the overruled cases and
their matches, we use a series of t-tests comparing means along the full length of each relevant curve.

**Results.** Table 1 provides comparisons of the overruled and matched cases. You can see that the two types of cases are comparable in both in and out degree means and distributions.

|TABLE 1|

Examining the overall depreciation figures for overruled cases, their matches, and for the entire population of Supreme Court decisions (Figure 1) demonstrates how cases that go on to be overruled are different from typical cases. While the rates of depreciation appear similar, overruled cases attract many more citations than the population as a whole. These cases are legally and socially important. After all, if they did not attract citations, there would be no need to overrule them.

|FIGURE 1|

At this scale the depreciation curves do not clearly show us whether overruled cases depreciate at a rate significantly different than their matches. Figure 2 helps address this by showing rescaled depreciation curves for both overruled and matched cases. You can see that overruled cases receive a significantly lower proportion of their total citations in the early part of their lives and a significantly higher proportion in their twilight years. Thus, the answer to RQ1 is that yes, overruled precedents *do* indeed have unique depreciation rates. They are more active citation targets later into their lives than similar cases.

|FIGURE 2|

**Authority scores.** The authority score graph (figure 3) shows overruled cases have consistently higher authority scores (i.e. are cited by more important cases) than their matches. However, this difference is only statistically significant when precedents are aged...
between 13 and 19 years. In response to RQ2 we find that overruled cases do indeed have
greater authority than similar cases. Decisions that go on to be overruled are cited by more
influential laws than cases that receive a similar number of citations, but do not go on to be
overruled.

[FIGURE 3]

*Hub scores.* The hub score graph (figure 4) demonstrates a similar trend as the authority
score graph. Overruled precedents have higher hub scores than their matches (i.e. cite more
important cases). The differences are statistically significant for precedents aged between 3
and 19 years. In response to RQ3 we find that overruled cases are more well grounded in the
law than cases that make a similar number of citations. Those cases that go on to be
overruled tend to cite cases that are more central and important than those cited by their
matches.

[FIGURE 4]

**Discussion.** Comparing overruled precedents and matched cases demonstrates that bad laws
are indeed significantly different from their peers not only in terms of their content, but also
in terms of the authority of cases they cite and in the way they themselves go on to be cited.

*Depreciation.* This study demonstrates that precedents that go on to be overruled are
originally less popular than their matches, but as the date of their eventual overruling
approaches they attract increased attention, essentially slowing their depreciation.

*Centrality.* Overruled cases not only have distinct life cycles, they also have higher authority
and hub scores than cases with similar citation profiles. Admittedly the differences are not
statistically significant across the entire age range, but this is partially a product of the
decrease in the number of case observations as the age increases (i.e. there are many fewer
Higher authority and hub scores for overruled cases suggest that they focus on areas of the law that are particularly important and influential. The causal direction between bad law and network centrality is not clear. It could be that bad laws become central as they attract attention and are eventually overruled. Alternately, it could be that their centrality lends them a prominence that leads judges to feel obliged to explicitly overrule them in the process of developing the law. Perhaps the most plausible explanation is that there is an interdependency between bad law and network centrality: laws that are bad attract attention and thus become more central within the network and at the same time network centrality brings attention to potentially bad laws making them more likely to be a target for overruling.

**Legal change.** On its surface, this study explores quantitative citation metric differences between bad laws and good laws. More deeply, it is an exploration of how legal change takes place, and when the legal system uses its most contentious and sudden implement of legal change. What we’ve learned about the lifecycle of overruled precedents provides some insight into why overrulings occur. Cases that are overruled tend to be firmly situated within a strong body of precedent (i.e. have high hub scores), exert above average influence on the legal system (i.e. have high authority scores), and remain relevant and contentious later into their lives than other cases (i.e. have slower depreciation curves). When legal change takes place, it takes place at the top. Active, central, authoritative and influential cases are more prone to being overruled than more peripheral cases.

**Conclusion.** We have learned that overruled precedents stand out not only from average Supreme Court decisions, but they are also more central than matched cases. Their distinct depreciation curves show that these cases remain relevant longer than their peers and
engender increasing contention in the years that lead up to their overruling. Ultimately, we see that foregoing the norms of *stare decisis* is a step reserved for the most pressing cases, when highly central precedents remain contentious long after they are originally decided. When this occurs, Supreme Court justices sometimes take the extreme step of overruling previous decisions in order to reduce these highly central disagreements and, by excising the offending point of contention, promote legal system stability.
### Tables and Figures

#### Table 1.

<table>
<thead>
<tr>
<th></th>
<th>Overruled cases</th>
<th>Depreciation</th>
<th>Out degree matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean time to overruling</td>
<td>32.46 years</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Mean in degree</td>
<td>17.25 (s.d. 18.90)</td>
<td>16.59 (s.d. 20.17)</td>
<td>10.77 (s.d. 15.96)</td>
</tr>
<tr>
<td>Mean out degree</td>
<td>10.05 (s.d. 11.17)</td>
<td>9.45 (s.d. 9.29)</td>
<td>9.74 (s.d. 10.02)</td>
</tr>
</tbody>
</table>

#### Figure 1

![Depreciation vs. Age](image)
Figure 4
References


Knowledge Network Based on Legal Issues

Paul Zhang
LexisNexis
Miamisburg, OH, USA
Paul.zhang@lexisnexis.com

Harry Silver
LexisNexis
Miamisburg, OH, USA
Harry.Silver@lexisnexis.com

Mark Wasson
LexisNexis
Miamisburg, OH, USA
Mark.Wasson@lexisnexis.com

David Steiner
LexisNexis
Miamisburg, OH, USA
David.Steiner@lexisnexis.com

Sanjay Sharma
LexisNexis
Miamisburg, OH, USA
Sanjay.Sharma@lexisnexis.com

Abstract
Most text-based Systems use Key Terms, Concepts, Entities, Topics, etc. as basic semantic Units to operate on. In this paper, we describe a new type of data derived from the case corpus - the Library of Legal Issues. We believe that the collection of Legal Issues, being data-driven and semantically specific, can be seen as a summary or a condensed version of the Legal Knowledge, and would support deeper analysis and study of legal principles.

Introduction
Research and work has been going on to represent and model cases and other legal materials, the goal being to allow for better analysis, understanding, or even inference on the legal knowledge embedded in the corpus. Among all the studies and research efforts, networks based on semantics of the legal corpus have been gaining more attention recently.

Work on Legal Citation Networks is motivated by some of the intrinsic characters of citations. Legal citations are highly semantic; they are closely related to legal or editorial decisions in the right contexts [12, 19]. This association makes possible the exploration and analysis of relations and other semantic aspects of legal documents. Citations are also asymmetric and evolve over time as new documents are created [9]. This means that the structure of legal citation network is comparatively stable; effects or changes caused by Time would add to the general structure of existing network gradually over time. And these changes may be indicative of shifting of the general legal domain.

In fact, researches were conducted to explore patterns or trends based on citation networks in order to gain insights into the time developing structure or evolutions of the law [2, 9]. Fowler [3] reported that, based on the complete network of U.S. Supreme Court data (1754 – 2002), they were able to calculate Authority Scores of Court precedents, which were confirmed by legal experts. More importantly, their analysis was able to reveal trends or patterns of some Supreme Court decisions, and at some points, to predict the “rise and fall” of power of precedents. Winkels [18] and Mazzega [11] reported their research of Citation Networks in the Continental Legal traditions, and also confirmed the benefits of Citation Networking in analyzing their respective Legal Systems despite the lower citation counts in their court data (compared to a Common Law System, like that of the U.S.).

Researchers have also been working on mapping concepts into semantics-based networks to gain better understanding of the subject domain. For example, Szumlanski [14] reported automatically building a semantic network of related concepts by mining Wikipedia texts. Toyota [16] proposed an approach to create a hierarchical network of Japanese Law, which would enable analysis into the relationships between different laws. Research was also reported in exploring structures and relationships of legal principles or legislations [15, 6].

One common practice in text-based processing is to map data into vectors, graphs, or other formats, which can be seen as modeling of the domain knowledge in the respective subject paradigm. The linguistic Units used to build the representation vary depending on the goals and specific needs of the research. Based on work reported, base Units generally range from Normalized Terms found in the documents (e.g. Bartoloni [1], Taks [15]), Key Terms (Ji-Lung [7]), Concepts (Szumlanski [14]), Paragraphs (Kumar [8]), texts related to Reason-for-Citation (Zhang [19]), to other linguistic Entities (Malik [4]).

In the following, we explain how Legal Issues, a derived set of semantic items from the case data, can also serve as Base Units to model Legal Knowledge in a given Legal System (e.g. the U.S. Law). We believe that this new type of metadata and the derived structures and functions would help us better analyze and understand the part of legal knowledge that is normally not explicit. We start with discussion of the difference between Legal Concepts and Legal Issues.

Legal Concepts vs. Legal Issues
(Terms defined here are ONLY relevant to this context/discussion)

- A Legal Issue (in our contexts) can be seen as a Statement of belief, opinion, a principle, etc. It usually contains one or more “Concepts” to be meaningful. For example, here is a Statement:

  “Thirteen-year-olds should not own a vehicle.”

It has at least three Concepts in it: “13-year-old”, “vehicle”, and “to own”; and the author or speaker states clearly an opinion, a belief, or a piece of law. When such a statement has legal implication, it is a Legal Issue. Here are examples of Legal Issues found in cases:
a. “An inference is not reasonable if it is based only on speculation.”

b. “To constitute the crime of robbery, however, the use of force must be motivated by an intent to steal.”

c. “a statute will not be given an interpretation in conflict with its clear purpose, and that general words used therein will be given a restricted meaning when reason and justice require it, rather than a literal meaning which would lead to an unjust and absurd consequence.”

d. “... the initial question to be decided in all cases in which a defendant complains of prosecutorial misconduct for the first time on appeal is whether a timely objection and admonition would have cured the harm.”

• Concepts, on the other hand, are building blocks of discussion or Issues. The Concept “vehicle”, for example, is used in all the following Legal Issues:

a. “A police officer may approach a stopped vehicle and inquire about an occupant's well-being without intruding on the Fourth Amendment.”

b. “In Nebraska, a vehicle can be a tool of the debtor's trade if the debtor uses it in connection with or to commute to work.”

c. “State law governs the issue of security interests in motor vehicles.”

d. “In Idaho, it is a felony to purport to sell or transfer a vehicle without delivering to the purchaser or transferee a certificate of title duly assigned to the purchaser.”

Here we see that a Concept can be used in discussion of different Issues. Legal Issues, on the other hand, are more specific, and can serve as stand-alone statements relevant to the legal expert’s discussion and argument.

Legal Issues vs. Legal Topics

A Topic, in our context, is analogous to a class/item in the catalogue of a library collection. Topics are semantics-based, and can be seen as “holders” of other contents, articles, documents, or Legal Issues, etc. The above four Issues, for example, belong to different Topics: (only top level of the Topic hierarchy are listed)

a. “A police officer may approach a stopped vehicle and inquire about an occupant's well-being without intruding on the Fourth Amendment.” ⇒ <Criminal Law & Procedure>

b. “In Nebraska, a vehicle can be a tool of the debtor's trade if the debtor uses it in connection with or to commute to work.” ⇒ <Bankruptcy Law>

c. “State law governs the issue of security interests in motor vehicles.” ⇒ <Contracts Law>

d. “In Idaho, it is a felony to purport to sell or transfer a vehicle without delivering to the purchaser or transferee a certificate of title duly assigned to the purchaser.” ⇒ <Transportation Law>

On the other hand, Issues under the same Topic may be thematically similar, but they may relate to different legal principles, and are used for different situation or discussion. For example, the extended Topic, “<Bankruptcy Law . . . Specific Exemptions>” has these two Legal Issues under it:

a. “In Nebraska, a vehicle can be a tool of the debtor’s trade if the debtor uses it in connection with or to commute to work.”

b. “To be eligible for the ‘automatic’ homestead exemption under Cal. Code Civ. Proc. § 704.720, at least one of the debtors must have resided continuously in the dwelling from the time the creditor’s lien attached until the forced judicial sale”,

which are used to express application of exemption rules in two different bankruptcy contexts.

Through the above discussion, we can see that Legal Issues, being full statements, can better represent the semantics of legal documents. While Concepts, Topics and other types of linguistic Units tell us what a legal discussion is generally ABOUT, Legal Issues tell us what the discussion is specifically SAYING.

Extraction of Legal Issues from Corpus

At any given time, there is a finite number of Legal Issues being discussed even though we do not know what that number is. These Issues form the body of knowledge of the Legal System, and represent principles of the Law. Yet, for Common Law Systems, this kind of knowledge is, to a large extent, embedded in case documents in the form of free texts. This lack of comprehensive compilation of all laws in the System (as different from codified laws in Continental Legal Traditions) imposes difficulties for legal professionals as well as Information Systems based on Computers. Here we report our efforts in bringing a solution to this kind of situation.

Through data-mining, important Issues can be extracted from a case law corpus, and stored in a repository, or a Legal Issue Library. The building of the Library relies on a data-mining process that collects Legal Issues in the corpus based on semantics-based network traversing. This traverse function is able to link citations related to a starting Legal Issue during a recursive search in the network space [ref. 19]. The basic process is depicted in Figures 1-5.

In a Caselaw database, legal cases are connected by citations. When Case A cited Case B, the author usually had a particular Legal Issue in mind. Figure 1 shows this situation, where cases are connected by citations, and colors of the links represent different Legal Issues behind the citations.
Discussions of important Legal Issues can be traced or chained together along citation lines and stored in a repository as metadata of the legal corpus (Figures 2-4). At each cycle of the process, one Legal Issue is extracted, and a new Issue category is created and stored in the Repository (together with all the Incidents, i.e. the individual citations).

There are several post-processing steps that handle merging, representation, indexing, etc. of the extracted Library Items. During the process, Instances of the same Legal Issue are “normalized”, and collapsed into the same Issue Identity with links to their original cases. This allows cases to be grouped under the same Issue ID even when the language of discussion is varied. For example, the following lines from different cases show this kind of variation:

- Robbery is “the felonious taking of personal property in the possession of another, from his person or immediate presence, and against his will, accomplished by means of force or fear.” The intent to steal must be formed either before or during the commission of the act of force.

- According to Green, under California law, the crime of robbery cannot be committed if the intent to steal is formed after the murder.

- Defendant testified that he had not thought about stealing any of Mullins’ property until after the assault was completed. If defendant had not harbored a larcenous intent before or during the assault, the taking was theft rather than robbery.

- No robbery occurs when the intent to steal is formed after the use of force.

- Defendant claims his various admissions go to the killing and not the robbery. Further, he argues there was no evidence showing he formed the intent to rob before he killed the victim.

- Defendant’s claim of insufficient evidence is premised on a misunderstanding of the immediate presence element of robbery. So long as defendant formed the intent to take the Brandts’ possessions before killing them, he was properly convicted of robbery.

Despite the variation in linguistic expression, these lines are clearly statements of the same Legal Issue regarding the nature of the intent required to support charge of robbery, and are duly stored under one Issue ID.

Each Legal Issue thus extracted can be considered a small piece of Law in the U.S. Legal System. The collection of all Issues can be seen as a summary or condensed version of the Legal Knowledge. In the remaining part of the paper, we discuss some implications of this new type of metadata.

(* More information of steps in this Process can be found in U.S. Patents [20, 21, 22, 23]. *)
Representing Cases with Normalized Legal Issues

In the Common Law tradition, cases are normally argued with points or issues that are supported by legal precedents. Attorneys use citations to establish authority of the precedents in support of their propositions. An experienced legal expert is usually able to follow the line of argument straight through the many citations embedded in it [10]. To some extent, the citations and Legal Issues behind them form an approximate skeleton of the Case.

With Legal Issue Library in the background, and important Issues normalized and indexed as described above, a new layer can be created underneath the case data, where each case is represented by Legal Issues it contains: (see Figure 6)

<table>
<thead>
<tr>
<th>Case ID</th>
<th>Legal Issues found in Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>CASE_00000001</td>
<td>LLI_000055; LLI_000321; LLI_990775; ...</td>
</tr>
<tr>
<td>CASE_00000002</td>
<td>LLI_000780; LLI_017543; LLI_100095; LLI_020344; ...</td>
</tr>
<tr>
<td>CASE_00000003</td>
<td>LLI_000055; LLI_000781; LLI_970850; LLI_032455; ...</td>
</tr>
</tbody>
</table>

(Figure 6)

This new data is an extra semantic structure superimposed onto the legal data. It will facilitate calculation of distance between cases in a new direction, i.e. based on Legal Issues they share. The new metadata may also open up a door for more efficient study of legal principles, how they are used in legal arguments, and what kind of relationships they have among themselves, etc. (Not the focus of this paper.)

Building and Analysis of Legal Issue Network

One way to model knowledge of a subject domain is to map the key Semantic Units into a graph representation, where nodes (vertices) represent the Units, edges the relations between them. Szumlanski [14], for example, constructed a Semantic Network based on over 3,000 frequently-occurring terms in Wikipedia data. Their System distinguishes between different senses of words, and builds a Concept Network where Concepts are the designated senses in the context. Their calculation of semantic relatedness relies on co-occurrence of terms at the sentence level. Their example graphs showed Semantic Relatedness as well as Semantic Similarities between Concepts, some of the relations being not very direct or obvious from the surface.

Taks [15] developed an approach to connect “Norms” (“the smallest units of legislation”) in Estonian Legislation. Being a Civil Law System, Estonian Legislation has a hierarchy of Legal Norms (Codes/Acts), which, due to historical and other reasons, shows “structure and [the] referential indeterminacy”. The research aimed at mapping the Nodes in the hierarchy onto a graph to reveal dependencies and purposes of Legal Norms. Their approach analyzes into the predicate structure of sentences, and uses verb clauses as the semantic form of the Norms. Similarity between Norms is calculated with co-occurrences of verbs and other components in the verb clauses. Figure 7 shows an example of their resulting (partial) graph, which shows the dependencies and strength of relations of Acts and Laws in Estonian Legislation.

As discussed earlier, the collection of Legal Issues can be seen as a condensed version of knowledge of the Legal System, where each item (an Issue) can be considered a small piece of Law. This is especially important for a Legal System that follows Common Law traditions since, here, substantial areas of law are not necessarily codified in the same manner one might find for Estonian or other Continental Legal Systems. Thus, for these Common Law Systems, our extracted Library of Legal Issues will serve as a particularly effective vehicle for study of legal principles and their interactions.

Like other semantic Units in legal data, Legal Issues are connected by citations and associated semantic elements. When they are used as basic operation Units to form Networks, much of the more profound legal knowledge that has not been explicit or easily seen will be revealed and become obvious. The following is an example based on partial processing of a small data, which includes a few Legal Issues centered around a starting Issue, once again, the “Motivation Element Required for Robbery” (CL_001):

- **CL_001**: In order to constitute robbery rather than theft, the act of force or intimidation must be motivated by the intent to steal; if the larcenous purpose does not arise until after the force has been used against the victim, there is no joint operation of act and intent necessary to constitute robbery.

- **CL_002**: A reviewing court must “review the whole record in the light most favorable to the judgment below to determine whether it discloses substantial evidence — that is, evidence which is reasonable, credible, and of solid value — such that a reasonable trier of fact could find the defendant guilty beyond a reasonable doubt.”

- **CL_9**: Prejudice is shown when there is a “reasonable probability that, but for counsel’s unprofessional errors, the result of the proceeding would have been different. A reasonable probability is a probability sufficient to undermine confidence in the outcome.”

- **CL_11**: The quantum of evidence the people must produce in order to satisfy the corpus delicti rule is quite modest; case law describes it as a slight or prima facie showing.

- **CL_12**: The intentional commission of the underlying felony is not only an essential element of the crime of first degree felony murder; it is the sole basis for holding the killing is murder in the first degree.

- **CL_17**: Robbery is defined as the “felonious taking of personal property in the possession of another, from his person or immediate presence, and against his will, accomplished by means of force or fear.”

- **CL_27**: Conduct by a prosecutor that does not render a criminal trial fundamentally unfair is prosecutorial misconduct under California law.
only if it involves the use of deceptive or reprehensible methods to attempt to persuade either the court or the jury.

- **CL_48**: The force or fear element of robbery may be directed either to the initial taking of the property or to its asportation. Thus, even when the intent to steal arises after the use of force or fear, the offense is robbery and not theft if force or fear was used to escape with the property.

- **CL_147**: The trial court has a sua sponte duty to instruct on lesser included offenses when the evidence raises a question as to whether all of the elements of the charged offense were present and there is evidence that would justify a conviction of such a lesser offense.

- **CL_196**: A defendant claiming ineffective assistance of counsel must first establish that "counsel's representation fell below an objective standard of reasonableness...[P]...under prevailing professional norms."

- **CL_213**: The trial court is required to instruct sua sponte only on general principles of law relevant to issues raised by the evidence and on particular defenses when a defendant appears to be relying on such defense and there is substantive evidence to support it.

- **CL_264**: An error in failing to instruct on lesser included offenses requires reversal unless it can be determined that the factual question posed by the omitted instruction was necessarily resolved adversely to the defendant under other, properly given instructions.

As noted, these disparate Issues form a small network, a part of the general Legal Issue Network of the U.S. Law, where Nodes (Issues) are linked by weighed Edges. (see Figure 8).

![Figure 8](image)

Not all members of the Network play equally strong roles in establishing network cohesion; and even within a network, sub-networks can be identified.

For example, from this Network, we see that "Definition of Robbery" (CL_17) has stronger connection to the Starting Issue (CL_001); it also has stronger connection to a few other Issues, e.g. "Review of Evidence" (CL_002), and "Use of Force" (CL_48). And the Issue, “Court’s duty to instruct on the lesser” (CL_147) has stronger connection to “required reversal or resolution when error is made with that respect” (CL_264).

Both Attorneys and Judges use Legal Issues in their arguments; the selection and use of these Issues influences, to a large extent, the outcome of cases. The Legal Issue metadata may provide a way to study into the logical thinking and strategy behind argument of Cases. Legal experts may also find it useful as to when and how Cases share the same set of Issues when formulating their respective argument strategies.

For example, based on the small Network above, we found that, among the small set, two cases showed particularly high overlap of "Issue usage". Specifically, “PEOPLE v. CANTWELL, 2004 Cal. App. Unpub. LEXIS 1833” and “People v. Frye, 18 Cal. 4th 894” used the following four Legal Issues (identified from the Network):

- **A reviewing court must “review the whole record in the light most favorable to the judgment below to determine whether it discloses substantial evidence -- that is, evidence which is reasonable, credible, and of solid value -- such that a reasonable trier of fact could find the defendant guilty beyond a reasonable doubt.” (CL_002, supra)

- **Robbery is defined as the “felonious taking of personal property in the possession of another, from his person or immediate presence, and against his will, accomplished by means of force or fear.” (CL_17, supra)

- **Conduct by a prosecutor that does not render a criminal trial fundamentally unfair is prosecutorial misconduct under California law only if it involves the use of deceptive or reprehensible methods to attempt to persuade either the court or the jury. (CL_27, supra)

- **A defendant claiming ineffective assistance of counsel must first establish that “counsel's representation fell below an objective standard of reasonableness...[P]...under prevailing professional norms.” (CL_196, supra)

Legal experts may find this higher degree of sharing of Issues as an indication of two cases sharing similar Factual Patterns, similar Arguments Strategies, or both.

### Conclusion and Future Work

We have reported our work on extraction and building of the Library of Legal Issues, and implications of this new type of metadata. We believe that such a collection can be seen as a summary or condensed version of legal knowledge found within the corpus of a known jurisdiction. The Library can serve as an added semantic layer for the legal corpus, and could serve as well as foundation for different semantics-based research tools. Our future research will continue to explore associations from Issues in the Library to other types of metadata, and address the potential of this newly derived metadata collection in supporting and creation of legal research and reasoning tools that could assist users in ways not yet addressed by existing research solutions.

### References


International Conference on Artificial Intelligence and Law. 2009.


Network Analysis of Manually-Encoded State Laws and Prospects for Automation

Patricia M. Sweeney, Elizabeth F. Bjerke, Margaret A. Potter, Hasan Guclu, Christopher R. Keane (Public Health Team)
Kevin D. Ashley, Matthias Grabmair, Rebecca Hwa (Machine Learning Team)

Abstract. This paper presents a methodology for applying network analysis to explore relationships within civil networks established by state law and to compare similarly-purposed legal systems across states. The methodology has been applied to state public health systems as defined in state laws for emergency preparedness. The paper describes a methodology for annotating the statutes, illustrates and explains the resulting Statutory Task Network Diagrams, describes analyses of the diagrams and their policy implications, and discusses prospects for automating the encoding using machine learning.

1. Introduction

Civil societies institute systems to achieve social goals in areas like public health, transportation, and communications. The public health system, for instance, comprises a network of institutional and individual actors, including state, federal, public and private agencies, whose purpose is to improve people’s health, in part through emergency preparedness and response.

Statutory law directs and coordinates the interactions in such a network. Although differing from state to state, in general, the statutes direct actors to take actions in coordination with other actors for the purpose of achieving goals within a particular timeframe and subject to certain conditions. At the University of Pittsburgh Graduate School of Public Health, policy analysts in the Public Health Team (PHTeam) have annotated statutes for those actors, goals, timeframes and conditions, represented them in network diagrams, and, using network analysis, studied the results to evaluate and compare how and how well states have prepared for public health emergencies. To ameliorate the need manually to annotate the statutes, the Machine Learning Team (MLTeam) has investigated if and how a program can be trained to automatically encode such statutes.¹

This paper explains the methodology for annotating the statutes (section 2), illustrates the Statutory Task Network Diagrams (section 3), describes analyses of the diagrams and their policy implications (section 4), and discusses prospects for automating the encoding (section 5).

2. Methodology for Manually Annotating Statutes

The PHTeam selected eleven states from which to collect relevant statutes to cover a variety of geographic and demographic conditions relating to public health emergencies. They include rural and urban states with different risk profiles. Kansas and Texas are in the US “Tornado Alley,” earthquakes regularly affect California and Alaska, and New York has experienced incidents of urban terrorism. The PHTeam also developed a standard list of public health system (PHS) agent types. (See Table 1.)

---

¹ We are grateful to the University of Pittsburgh’s University Research Council Multidisciplinary Small Grant Program for funding this project. This publication was also supported in part by the Cooperative Agreement 5P01TP000304 from the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the CDC.
To construct a statute text corpus, the PHTeam drafted LEXIS queries for PHS agents (see Table 1) and relevant statutes with query strings like: “Disaster”, “Emergenc!”, “Hazard!”, “Preparedness”, “Terror!”, “Infec!”, “Communicable”, “Pandemic”, “Epidemic”, “Food borne”, “Emergency or Disaster w/p Nuclear or Radi!”. Searches of the states’ statutory databases yielded 33,056 statutes.

<table>
<thead>
<tr>
<th>PHS Agent</th>
<th>&quot;Search Term&quot; and/or Definition</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Public Health Agents</td>
<td>Inclusive term used in a legal directive for “all” or “any” public health system agent</td>
<td>ALLPHAgents</td>
</tr>
<tr>
<td>Administrative Agencies</td>
<td>A department of government, not legislative or judicial, with administrative power delegated a legislature</td>
<td>Admin Agencies</td>
</tr>
<tr>
<td>Community Health Center</td>
<td>“community health”</td>
<td>CHC</td>
</tr>
<tr>
<td>Emergency Medical Service</td>
<td>&quot;EMS&quot; or &quot;emergency management system&quot; or &quot;emergency medical service&quot;</td>
<td>EMS</td>
</tr>
<tr>
<td>Elector Official</td>
<td>&quot;governor&quot; or &quot;mayor&quot; or &quot;political subdivision&quot; or &quot;county commissioner&quot; or &quot;county executive&quot;</td>
<td>Elector Officials</td>
</tr>
<tr>
<td>Emergency Management Agency</td>
<td>“emergency management agency”</td>
<td>EmerMgmt</td>
</tr>
<tr>
<td>Employer</td>
<td>&quot;employer&quot;</td>
<td>Employer</td>
</tr>
<tr>
<td>Faith Institution</td>
<td>“faith or relig! or church”</td>
<td>FaithInst</td>
</tr>
<tr>
<td>Governmental Public Health</td>
<td>&quot;board of health&quot; or &quot;health department&quot; or &quot;department of health&quot;</td>
<td>GovPublicHealth</td>
</tr>
<tr>
<td>Hospital</td>
<td>&quot;hospital”</td>
<td>Hospital</td>
</tr>
<tr>
<td>Law Enforcement</td>
<td>“law enforcement” or “police”</td>
<td>Law-Enforcement</td>
</tr>
<tr>
<td>None</td>
<td>[no agent named in directive]</td>
<td>None</td>
</tr>
<tr>
<td>Other</td>
<td>An entity not otherwise defined</td>
<td>Other</td>
</tr>
<tr>
<td>Public/Individuals</td>
<td>Individuals or groups not subordinate, functionally related, or connected to a governmental agency or another public health system agents</td>
<td>Public</td>
</tr>
<tr>
<td>Special Response Teams</td>
<td>A complement of individuals or entities organized to provide a specified aspect of emergency preparedness, planning or response according to standards developed by an emergency management agency or department of homeland security.</td>
<td>SpecRespTeam</td>
</tr>
</tbody>
</table>

After a screening process, 4,917 relevant statutory provisions from the eleven states remained, where a *statutory provision* is a statutory element identified by a root-level §-citation (e.g., 4 Pa. Code §3.24). The PHTeam developed a coding scheme, comprising nine categories (see Table 2), to capture information about *which actors*, a statutory provision directs, take *what actions* in coordination with *which other actors* for the purpose of achieving *what goals* within *what timeframe* and subject to *what conditions*. Each category has a list of alternative values. For instance, there are 29 PHS agent types. The coding scheme is defined in a comprehensive codebook.²

At various times, from 4 to 8 human coders manually encoded the statutory provisions in a table, where each provision corresponded to one or more lines. Each line explicitly referred to a statutory provision and recorded a code label for each of the nine categories. Each provision had one line for each task it directed, since some provisions directed more than one task. The 4,917 relevant provisions yielded 17,163 lines of codes in the coding table.

The PHTeam evaluated inter-rater reliability of the coding. An approach recommended in [10; 11] was used to compute a sample size sufficient to detect a relative error rate as low as 20%. A sample of 2% of the statutory provisions was drawn from the first 5 states, with one randomly selected variable per provision. An external analyst reviewed and re-coded the sample variables [13]. The inter-rater reliability was 68.8% (95% CI: 63.1%, 74.9%).

3. Statutory Task Network Diagrams

The PHTeam constructed a series of statutory task network diagrams from the table to summarize and display information about how a state’s statutes direct certain tasks for certain purposes. A "Statutory Task Network Diagram" (STN Diagram) is a network diagram whose nodes represent statutorily-directed agents and each of whose directed edges represents a task that a statutory provision directs one agent to perform with respect to another agent. Since each task could involve an agent in different actions, goals, purposes, emergency types, and conditions, the PHTeam focused on what would be a fruitful basis for comparing different states’ legal frameworks. For each state, STN diagrams were generated to represent the tasks by purpose. A diagram was created for each of three relevant purposes: preparedness, response, and recovery, and a fourth diagram represented tasks for any and all purposes. Additional diagrams were created for more specialized purposes, for instance, infectious disease surveillance, a combination of actions and goals with infectious disease as the type of emergency.

For instance, Figure 1(a) shows the STN diagram for Pennsylvania (PA) statutory provisions dealing with public health emergencies that direct PHS agents to engage in tasks for the purpose of preparedness. The list at the left of the diagram shows the PHS agents that PA law leaves out with respect to preparedness, which is one example of useful information in an STN.

In analyzing network diagrams, a key concern is defining measures that correspond meaningfully to domain phenomena of interest. The PHTeam identified four such measures: density, inclusiveness, degree, and strength [3; 8; 19; 20; 21].

<table>
<thead>
<tr>
<th>Measure</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density</td>
<td>Total number of connections in a graph, expressed as a proportion of the maximum possible number of connections: ( n(n-1)/2 ) where ( n ) is the number of agents.</td>
</tr>
<tr>
<td>Inclusiveness</td>
<td>Number of agents with network connections, that is, the number of possible agents minus the number of unconnected agents.</td>
</tr>
<tr>
<td>Degree</td>
<td>Degree of connection of an agent is the number of connections it has.</td>
</tr>
<tr>
<td>Strength</td>
<td>In a weighted network, strength is the total weight of incoming (“in-strength”) and outgoing (“out-strength”) edge weights for the connections of each agent.</td>
</tr>
</tbody>
</table>

These measures can be interpreted in ways that bear on the design of the public health systems. The density of a statutory task network like that in Figure 1 (a) measures how “spread out” the legally mandated tasks are over all agent types in the network. Inclusiveness of a statutory task network measures the degree of agent-type involvement in the task.
Since an STN Diagram like that in Figure 1 (a) is a directed network, its degree has two components: an agent’s in-degree, the number of connections directed to it, and its out-degree, the number of connections directed from it. In a weighted (valued) network, weights can be assigned to the edges. The PHTeam used the frequency of pairings between agent types as a measure of an edge’s weight. As noted in Table 3, a weighted network supports a measurement of strength, that is, the total weight of incoming (“in-strength”) and outgoing (“out-strength”) edges for the connections of each agent type. In Figure 1 (a), an agent-node’s out-strength is a measure of its outgoing load of directives; the agent-node’s in-strength is its incoming load of directives from other agent types.

4. Analysis with Task Network Diagrams

The analysis suggests hypotheses worth investigating through detailed investigation of the laws’ substantive content or comparison with expectations in public health literature. For instance:
1. In all 11 states, the network density is lower for purposes of recovery than for other purposes. The recovery-purpose laws specify roles for a lower proportion of PHS agents.
2. 10 of 11 states’ preparedness networks are more inclusive than the networks for the recovery phase.
3. The average degree for PHS agents (i.e., number of connections per agent) for “all” purposes varies widely among states: from a high in NY (13.03) to a low in ND (6.03). That is, states differ in their attention to defining legal inter-relationships among PHS agents.
4. Across all states, the average degree among PHS agents is highest for the response phase in all but two states (NY and WI) and lowest for the recovery phase in all states.
5. Across all states, average strength of PHS agents (including both in-strength and out-strength) is greatest for the response purpose and least for the recovery purpose. The wide range of average
strength of PHS agents for the response purpose across states suggest that states’ approaches to allocating response tasks vary substantially.

The network analyses enable visualizing complex legal relationships and comparing legal systems. The STN diagrams can dramatically present some comparisons like those mentioned. For instance, Figures 1 (b-d) are diagrams of laws of New York (NY), Pennsylvania (PA) and Alaska (AK) directing PHS agents for the purpose of infectious disease surveillance. The differences are apparent in the degree and strength of connections in the networks representing the three states’ laws set up to address the same task. While assessing the suitability of a given state’s laws requires some independent measure of quality, the contrasting diagrams suggest the utility of searching for one.

Policy makers can identify aspects of legal systems that may need to be strengthened. The low density and inclusiveness of the networks representing emergency recovery laws, for example, suggest an area for legislative change to improve community resilience by identifying and directing additional PHS agent types. Comparing states with shared geographic or demographic attributes could be especially useful: for example, Maryland’s high-strength of PHS agents in recovery laws could have implications for nearby Pennsylvania or in another coastal state such as Rhode Island.

The network measures of emergency laws can also reveal important dependencies. Elected officials’ and administrative agencies’ consistently high strength for emergency preparedness, response, and recovery indicates that their incapacitation in an emergency could hobble an entire PHS. States could redress this by creating redundancy for these participants’ roles.

Of course, the STNs do not capture much of the laws’ semantic content. At best they enable a visual exploration that complements traditional statutory analysis. In addition, the de jure directives of legal statutes do not necessarily correspond to the de facto realities of PHS agents working on the ground. Moreover, the statutes are subject to change. Also, constructing STN diagrams depends on manual encoding of a great deal of statutory language. We now turn to prospects for automating that encoding.

5. Prospects for Automated Encoding

The MLTeam has developed a framework to encode statutes automatically from plain text using classifiers trained on data from the same state (Pennsylvania). The encoding task was decomposed into several classification problems. Preliminary results showed that most of the learned classifiers performed better than baselines, although the results fell far short of a practical solution [9].

The statutory provisions were segmented into a tree structure of clauses with leaf nodes as sentences and divided into chunks. A chunk is a subpart of a provision comprising all the clauses (and their associated text) from tree root to a specific clause node and is identified by a unique citation. Each chunk was translated into a feature vector using bag-of-words (i.e., an unordered collection of words) and TFIDF (Term Frequency / Inverse Document Frequency) representations, where a word’s feature value becomes a numerical measure of its relative importance.

The MLTeam trained a model to predict whether a chunk satisfies the criteria for coding for purposes of the public health analysis. They also trained models to predict (an unspecified number of) labels for a chunk’s feature vector. These labels were the human-assigned codes as per Table 2. As one can see from the third column, the number of possible target codes varies greatly in size, inhibiting effective prediction. The MLTeam compared the classifiers to two baselines that determine the single most frequent code for a given attribute from the training data and predict it for all the validation data with or without the use of a small dictionary used in the manual coding.

In the binary relevance classification, ML techniques improved precision over the baselines, but at the cost of lower recall. With respect to the single-label prediction, the ML approach significantly reduced classification error. For most multi-label classifiers (acting agent, goal, emergency type, condition and time frame), the ML prediction outperformed the baselines but not for receiving agent and purpose. Predicting an unspecified number of codes for attributes of interest and discriminating among ranked codes to achieve good precision remain as major challenges. Scarcity
and dominant codes in the data caused as yet unresolved problems [9].

6. Related Work

Network analysis can characterize network features, analyze relationship patterns, and supports graphical comparisons and presentation of results. Previous studies have applied network analysis to communication patterns of workers within local health departments [17], to word co-occurrence and citation patterns in the United States Code [2], to citation patterns in French legal codes [15], and to interpreting email in e-discovery [12]. As far as known, no prior research has applied network analysis to U.S. public health laws or to the relationships among agents in a PHS.

A substantial amount of published research addresses automatically classifying statutory texts in terms of: major types, abstract categories, and subject matters (e.g., "Administrative Law", "Intellectual Property Law") [7; 18; 16]; norm type (e.g., definition, permission, obligation) [14]; norm features (e.g., duty, duty bearer, action, object of action) [1; 6; 4; 5]; and regulatory function (i.e., functional information about what entities and events a statute governs, for what purpose, and subject to what constraints) [9], the approach used here.

7. Conclusions

In this work, researchers in public health and in artificial intelligence have developed methodologies for applying network analysis to explore relationships within civil networks established by state law and to compare similarly-purposed legal systems across states, and explored techniques aimed at automating the manual encoding of statutes required to construct Statutory Task Network Diagrams. The methodologies have been applied to, and the resulting network analyses have policy implications for, state public health systems as defined in state laws for emergency preparedness.

8. References

Figure 1. Statutory Task Network Diagram of emergency laws directing PHS agents for the purpose of: (a) Preparedness in Pennsylvania, and of Infectious Disease Surveillance in (b) New York (NY), (c) Pennsylvania (PA) and (d) Alaska (AK). Colors denote if legal directives are bi-directional between two agents (red) or uni-directional from one agent to another (blue). Weights of lines indicate frequency of legal directives between agent pair.
From structure to function: exploring the use of text and social network analysis in criminal investigations

NICOLA LETTIERI, DELFINA MALANDRINO, CARLO RINALDI, RAFFAELE SPINELLI*


1 INTRODUCTION

This paper presents an ongoing research on the applications of computational tools in the analysis of structural and functional features of criminal organizations involving both researchers (legal and computer scientists) and public prosecutors. Drawing inspiration by a sociological study using network analysis techniques to compare features of two criminal organizations belonging to Sicily's Cosa Nostra and Campania's Camorra, the research aims at developing a framework, named CrimeMiner, that combines information extraction, network analysis and visualization methods to support investigation and the fight against criminal organizations. After a brief introduction to the applications of Social Network Analysis (SNA) in criminal investigation, the paper offers an overview of the results so far achieved from a technical and methodological point of view sketching future developments that appear to be challenging for both Criminology and Legal Informatics.

2 SOCIAL NETWORK ANALYSIS AND CRIMINAL INVESTIGATION

Among the various application contexts of SNA that are arousing the interest of legal scientists and practitioners, one of the most promising is the study of the social underpinnings of legally relevant phenomena, or, more in detail, the analysis of criminal groups and networks from a structural and functional point of view.

The social dimension plays a crucial role in the evolution of crime: a large part of criminal phenomena from drug networks1 to international terrorism2;

---

* Nicola Lettieri, researcher at ISFOL, Rome, is adjunct professor of Legal informatics at the University of Sannio and of Computational Social Sciences at the Department of Computer Science of the University of Salerno. Delfina Malandrino is assistant professor of Computer Science at the University of Salerno. Carlo Rinaldi is deputy prosecutor, at the Public Prosecutor's Office of Sala Consilina (Salerno). Raffaele Spinelli is a PhD student in Computer Science at the University of Salerno.


from pornography trafficking to hacking and other cybercrimes, is strongly conditioned (inhibited or facilitated) by relational dynamics. Criminals are not isolated: they are nested within communities, drawing support from members of the community. In this context, the SNA approach, thanks to the potentialities shown in mapping and measuring the social landscape, is becoming more and more interesting.

Nowadays, the achievement of results by SNA techniques is facilitated by the boundless amount of digital by-products (e-mails, mobile phone calls, credit-card operations, Internet searches, interactions mediated by social networks) generated by social life in the modern world: the digital data-stream and advances in technology are pushing both researchers and law enforcement agencies to figure out new tools and methodologies to illuminate the structures and dynamics of criminal networks.

The idea has recently gained growing popularity: over the past 10 years there has been an increasing number of experiences in the development and in the use of SNA-based investigative software. Criminal Network Analysis (CNA) is a flourishing research area in which criminology, organized crime research, social network theory and computer science converge with other disciplines to offer new insights into crime by means of innovative data mining tools and applications working to unveil hidden patterns in large volumes of crime data and investigative documents.

3. BUILDING A FRAMEWORK FOR CRIMINAL INVESTIGATION

Taking a cue from the scenario so far sketched, we decided to explore the application of information extraction and SNA techniques in criminal investigation. In this vein, fundamental inputs came from the collaboration with Italian deputy prosecutors that highlighted two circumstances relevant for our research:

If we put aside more “traditional” information systems (databases containing complaints, criminal records, police reports, etc.), criminal investigation remains primarily a manual process although with significant differences depending on the context taken into consideration. Computational tools most frequently used by public prosecutors do not provide either advanced information extraction functionalities or structural analysis of network knowledge from criminal justice documents. Furthermore, investigative information, even when in digital format, is very often unstructured.

Tools allowing to gather, analyze, visualize and extract information from investigative documents produced by the prosecutor’s office and judicial police would be considered extremely useful. Particular attention has been aroused by visual and “scientifically grounded” tools illuminating not only the extension, the topology and the patterns of groups of people involved in illegal activities, but also the role of specific individuals in the organization. After all, knowledge about the structural and functional properties of criminal networks is fundamental for both investigation and the development of effective prevention strategies.

Starting from the consideration of these elements, we have initiated a project in which domain experts (lawyers, judges) collaborate with computer scientists to develop and test, in a real investigative scenario, an integrated framework, named CrimeMiner, that provides functionalities to:

- enable the marking up and the automatic extraction of information necessary for performing network analysis of criminal groups;
- generate and analyzing graphs of criminal networks under investigation;
- provide a diachronic view of the evolution of criminal groups;
- organize and visually manage investigative material;
- make predictions about the potential of an individual to belong to a criminal group

According to the claims of the public prosecutors involved in the project, an environment of this kind not only could offer insights into criminal phenomena difficult to obtain in any other way, but could also have an impact in the dynamics of the investigation and trial. The graphs and the results of analysis carried out on them could be a tool to facilitate information sharing and collaboration between the judges and the judicial police; moreover, they could be used to support requests for committal for trial with the evidence of a scientific kind with particular value for the argumentation.

Our research has developed so far in two different phases:

**Phase 1**

It focused on the task of becoming familiar with the parsing and the analysis of legal proceeding by applying SNA techniques on telephone tapping.

From a practical point of view, the starting point of the activity was offered by an interesting research\(^9\) of a sociologist of deviance, Attilio Scaglione, who not...
only brought us closer to the use of SNA in the study of criminal phenomena but also aided us in the retrieval of documents on which to start experimenting. Oriented towards purely scientific purposes, the work of Scaglione analyzes the characteristics of the social network emerging from telephone tapping reported in two remand documents\textsuperscript{10}. Using one of these two documents, we started by posing as first objective the replication of his findings, by developing a specific tool whose description is presented in the next Section.

Phase 2
In the light of the results obtained, and taking cue from the discussions with judges actively engaged in investigation activities, we started to work for:

a) testing the reliability and the adaptability of the tool built during the previous phase, using it to analyze other remand documents.

b) extending the set of the information to gather and analyze for the construction of the network (not just the telephone tapping but also dates, places, criminal records, etc.)

c) creating a tool allowing to collect and mark up \textit{ab initio} with the necessary metadata information to be used for building graphs.

4. Phase 1: Information Extraction and Structural Analysis

In this phase we have focused our attention on the extraction of information from a particular type of pleading (request for pre-trial detention, request for an arrest warrant) and the consistent application of SNA metrics to this information. Our case study was represented by one of the remand documents analyzed by Scaglione, whose main properties are shown in Table 1.

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crime class</td>
<td>Organized Crime (Camorra)</td>
</tr>
<tr>
<td>Analized documents</td>
<td>Remand document</td>
</tr>
<tr>
<td>People involved</td>
<td>73</td>
</tr>
<tr>
<td>Location</td>
<td>Quindici, Avellino, Italy</td>
</tr>
<tr>
<td>Pages of the document</td>
<td>3000</td>
</tr>
<tr>
<td>Number of telephone and wiretap</td>
<td>2791</td>
</tr>
<tr>
<td>Number of phone under tapping</td>
<td>300</td>
</tr>
</tbody>
</table>

\textit{Table 1- Properties of the first case study}

Before starting with details about the implementation, we have to emphasize that an obstacle in this area is the difficulty of finding scanned documents, or other documents that are readily convertible into formats that can then be adapted for automatic analysis. As anticipated, we have analyzed a remand document used by Scaglione containing information about nearly 70 persons belonging the \textit{Cava} family, a criminal organization from Quindici, Avellino

\textsuperscript{10} Data used for the construction of the graph were manually extracted by Scaglione from unstructured documents.
Italy). To represent the crime network studied by Scaglione we built a graph taking into account the following information:

- telephone tapping (nodes are connected if there was a communication)
- phone call direction (who calls whom)

The phone call direction could be interesting for deriving the importance of members of the organizations. We did not consider the content of the call and its semantic as our main and initial interest was in the structure of the resultant network.

Firstly, we needed to apply an initial “Text Filtering” step to make official documents suitable for the subsequent automatic “Parsing and Graph Generation” step. In this second step, the remand documents were automatically processed to analyze behaviors and relationships in order to produce, in the last step, namely, “Graph Visualization”, the corresponding visual representation. We will describe each of them in the following sections.

4.1 Text Filtering

During this step we filtered out unnecessary information and marked some specific parts needful for the network analysis. To this end, we have developed a simple parser (in Perl Language) that extracts entities (name, surname, telephone number) relevant for the graph. It was not possible to completely process the documents automatically because of their nature. In fact, they are known to suffer from the following critical deficiencies:

- Incompleteness: sometimes some information about nodes and relations are missing, making impossible to construct a global vision of the network that resembles as much as possible the real one. Examples include the minimization of the criminals’ interactions to avoid attracting police attention and the hiding of their interactions behind various illicit activities.
- Incorrectness: Incorrect data regarding criminals’ identities, physical characteristics, and addresses may result either from unintentional data entry errors or from intentional deception by criminals.
- Inconsistency: Information about a criminal who has multiple police contacts may be entered into law enforcement databases multiple times. These records are not necessarily consistent, involving inaccuracies when building the network.

4.2 Parsing and Graph Generation

During this step we find out the relations between the members of the organization. To do that, the parser firstly extracts from the remand documents (in a specific section) the list of people involved in the investigation that will represent the nodes of the network. It produces an XML file representing the

---

network, compliant with the GEXF file format specifications\(^\text{12}\), a language for describing complex networks structures, their associated data and dynamics.

### 4.3 Visualization

This last step allowed to visualize the resultant graph. We loaded the produced GEXF file into a framework for graph analysis and visualization named Gephi\(^\text{13}\), a framework that incorporates and offers a wide range of algorithms from the Graph Theory literature (i.e., algorithms for: metrics and statistical properties extraction; group discovery; visualization).

Once loaded the graph, we computed some statistics, such as, the importance of the organizations’ members. The graph representing our case study is shown in Figure 1, with nodes representing members (we used Node IDs instead of real full names). Color and size give information about group membership and relevance, respectively.

![Criminal network graph](image)

**Figure 1 - Criminal network graph (size proportional to Degree Centrality)**

Communities have been detected using the algorithms described by Blondel\(^\text{14}\). The size of the nodes is proportional to the degree centrality number of incoming and outgoing phone calls).

---

\(^{12}\) See: http://gexf.net/format


For each node we calculated degree centrality that allowed us to identify the more influential actors in the criminal organization considered in the remand document (Node2 and Node5 in Fig. 2), a result that reflects both findings manually obtained by Scaglione and trial evidences. More details about the metrics of the resultant graph are summarized in Table 2.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of vertices</td>
<td>76</td>
</tr>
<tr>
<td>Number of edges</td>
<td>176</td>
</tr>
<tr>
<td>Density</td>
<td>0.067</td>
</tr>
<tr>
<td>Diameter</td>
<td>4</td>
</tr>
<tr>
<td>Mean degree</td>
<td>5.117</td>
</tr>
<tr>
<td>Modularity</td>
<td>0.375</td>
</tr>
</tbody>
</table>

Table 2 - Metrics of the first case study

An interesting feature of the tool so far developed is the possibility to switch from a static to a dynamic version of the graph (see Figure 2). Taking into account the date associated to every telephone tapping, the tool is able to generate different overviews of the organization under investigation allowing researchers to examine the evolution of the criminal group over time, identifying temporal patterns and trends (growth, decline of criminal groups).

Figure 2 - Evolution of a criminal network over time

5. Phase 2: Analysis enhancement and tool features extension

In this second phase, we started testing the reliability and the adaptability of our tool analyzing new remand documents (and a new case study), related to drug
trafficking activities in the south of Italy. The properties of this new documents are shown in Table 3.

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crime class</td>
<td>Drug Trafficking Network</td>
</tr>
<tr>
<td>Analized documents</td>
<td>Remand document</td>
</tr>
<tr>
<td>People involved</td>
<td>19</td>
</tr>
<tr>
<td>Location</td>
<td>Sala Consilina, Salerno (Italy)</td>
</tr>
<tr>
<td>Pages of the document</td>
<td>476</td>
</tr>
<tr>
<td>Number of telephone and</td>
<td>480</td>
</tr>
<tr>
<td>wiretap</td>
<td>84</td>
</tr>
</tbody>
</table>

Table 3 - Properties of the second case study

According to the suggestions provided by hands-on investigation experience, we extended the set of data taken into account by the tool in order to include new information such as dates of telephone calls, residence of individuals, criminal records and so on. We experimented then filtering capabilities to allow prosecutors and investigators to identify, in the network under investigation, individuals showing given features (specific criminal records etc.) that can be considered circumstantial evidences of a criminal activity. The effect of applying a filter is shown in Figure 3 (filtering by city of residence).

![Figure 3 - Screenshot of Gephi interface by enabling filtering capabilities](image)

We also started to work on the implementation of a more comprehensive framework able to support the entire investigation, from the activities of the judicial police to those of public prosecutors. The design of the framework, that is going to be tested in real investigation, relies on a pipeline with two steps:
**Step 1**
Design and implementation of an editing tool enhancing the drawing up of documents (i.e., remand documents) allowing to collect and mark up *ab initio* all the necessary metadata information to be used for building graphs. The editing module will support investigators in reporting information in an already structured form. The idea is to overcome the problem of information extraction we experienced, allowing a fixed structure for any produced document. The framework will also allow judicial police and public prosecutors to easily share investigative information and to collaborate in writing documents reducing the burden of time consuming tasks. We plan to implement this module using an *ad-hoc* editing tool based on *Etherpad*\(^5\). Basic (and extendable) templates to insert relevant information (people under investigation, criminal records, data about telephone tapping etc.) will be available.

**Step 2**
Integration of tools and creation of facilities in order to offer users (both public prosecutors and judicial police) a unified workflow. We plan to make the environment available through Web interface. A specific attention will be paid, in this vein, to privacy and security issues, on both an infrastructure and application level.

### 6. Conclusions

Although in progress, experiments have given some preliminary outcomes that deserve to be briefly described. The most tangible results are found on the level of implementation. The work done so far has allowed us to explore the issues related to the parsing of unstructured investigative and prosecution documents and to obtain useful information on the processing of these kinds of documents. The analysis carried out are encouraging in terms of reliability in extracting relevant information: data derived from automatic analysis and the subsequent processing of the documents fit with trial evidences both in the first and the second case study. To all this, always on the application level, there is also the implementation timeline for the diachronic view of the criminal network.

From the point of view of the analysis of the phenomenon being investigated, the research suffers from a number of limitations arising primarily from the characteristics of the documentary material used: if a first limit is represented by the fact of having used so far few documents, another aspect that is not secondary is the nature of the information processed. Just for instance, the number and the direction of phone calls used for the construction of the graph, by itself, is not sufficiently and univocally meaningful: depending on the group, the region and the culture considered, a person playing a secondary role in a criminal organization can be the recipient of a large number of phone calls. The issue that came out during the discussion with the prosecutors emphasized how that which turns out to be relevant on the scientific level of a sociological-type study, albeit inspiring, often may not be significant for investigation purposes.

According to this consideration, it clearly emerged how the creation of a support tool for carrying out an investigation requires not only increasing the

\(^5\) http://etherpad.org/
amount and variety of information processed to build up graphs (e.g., gathering data from criminal justice databases) but also refining criteria used in processing judicial and police information trying to “embed” the inquirer’s know-how in information extraction procedures. In this perspective, it will be useful to use reliable methods to weigh the impact that single investigative information (e.g. a given type of criminal record) have on the probability that an individual is part of a group or a criminal organization.

The most interesting result, however, is found on the methodological level and the perspectives opened up by an interdisciplinary collaboration regarding crime analysis technologies are showing that they can have an impact on the development of interdiction and law enforcement strategies.

The research will continue in two directions devoting itself, on the one hand, to a deepening of the scientific issues linked to social and criminal network analysis and, on the other, to the design and implementation of new tools for the collection and analysis of investigative data.

The road ahead seems to be challenging both, on the one hand, for Criminology and Criminal law, and, on the other hand, for Legal informatics. For Criminology and Criminal Law, information technologies and SNA techniques represent an opportunity for methodological enrichment through an accentuation of the quantitative dimension of the study of crime. For Legal Informatics, the topic not only represents a new application context but also the opportunity to build on and improve the knowledge produced in processing information and legal documentation.

In the future, there is the possibility of integration with other computational social science methods, for example, with social simulations that may serve to combine SNA with the generative perspective of simulations to which the law has also begun to approach. The study of crime and, eventually, the prediction of its evolution can be based not only on the techniques of inferential statistics applied to large amounts of data but also on methods that take better account of the cognitive, social and institutional phenomena underlying criminal phenomena.

ACKNOWLEDGEMENTS
Authors would like to thank for their useful contribution, suggestions and comments dr. Attilio Scaglione from the University of Palermo and dr. Luigi Landolfi, deputy prosecutor of the Antimafia District Department of Naples.

---

16 See, for example, the special issue dedicated to “Simulation, Norms and Laws”, of the journal Artificial Intelligence and Law (Artificial Intelligence and Law, Vol. 20, No. 4, 2012 and Vol. 21, No. 1, 2013).

17 A particularly suggestive example of this technique is provided by PredPol (http://www.predpol.com/), a software environment for “predictive policing” developed under the guidance of anthropologist, P.J. Brantingham of UCLA.

A Network Analysis of Dutch Regulations
Using the MetaLex Document Server

Rinke Hoekstra¹,²

¹ Department of Computer Science / The Network Institute
VU University Amsterdam
rinke.hoekstra@vu.nl,
² Leibniz Center for Law, Faculty of Law, University of Amsterdam
hoekstra@uva.nl

Abstract. In this paper we explore the possibilities of using the Linked Data representation of all Dutch regulations stored in the MetaLex Document Server for the purposes of network analysis over the citation graph between regulations, both at the document level, and at the article level. We show that this is possible using relatively straightforward SPARQL queries, and present preliminary results of the analysis.

Keywords: Network Analysis, Linked Data, CEN MetaLex, Legislation, Regulation

1 Introduction

The MetaLex Document Server [2]³ (MDS) hosts all Dutch legislation and treaties available from spring 2011 onwards. The server was developed to overcome the limitations of the publicly available legislation published through the Wetten.nl portal of the Dutch government. Regulations available on MDS are published in two formats, CEN MetaLex, a flexible and jurisdiction agnostic XML format for publishing legal sources, and as Linked Data in RDF format. The Linked Data is browsable through a Pubby⁴ interface that operates on top of a 4Store SPARQL endpoint. The Linked Data format is highly suitable for graph analysis as the format itself (RDF) forms a graph of interconnected Web resources (URIs). At the time of this analysis (April 22, 2013), the MDS hosted 280,394,1322 triples across 33,643 document versions.

In this paper, we show how we can run tailored network analyses of the Dutch regulations published in MDS using relatively straightforward SPARQL queries. We present preliminary results of these analyses, and discuss the benefits of using Linked Data as source for network analysis. The networks and analyses discussed in this paper are published separately on Figshare.com [3].

³ See also http://doc.metalex.eu
⁴ See http://github.com/cygri/pubby.
Table 1. Network properties

<table>
<thead>
<tr>
<th>Measure</th>
<th>Document</th>
<th>Article</th>
<th>Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of nodes</td>
<td>14935</td>
<td>64018</td>
<td>4.286</td>
</tr>
<tr>
<td>Number of edges</td>
<td>33819</td>
<td>80082</td>
<td>2.368</td>
</tr>
<tr>
<td>Average degree</td>
<td>2.264</td>
<td>1.251</td>
<td>0.553</td>
</tr>
<tr>
<td>Avg. Weighted degree</td>
<td>9.117</td>
<td>3.749</td>
<td>0.411</td>
</tr>
<tr>
<td>Network diameter</td>
<td>16</td>
<td>8</td>
<td>0.5</td>
</tr>
<tr>
<td>Average path length</td>
<td>5.479</td>
<td>1.316</td>
<td>0.240</td>
</tr>
<tr>
<td>Avg. Clustering Coefficient</td>
<td>0.09</td>
<td>0.0021</td>
<td>0.023</td>
</tr>
<tr>
<td>Connected Components</td>
<td>492</td>
<td>7262</td>
<td>14.76</td>
</tr>
<tr>
<td>Number of SCC’s</td>
<td>14019</td>
<td>63303</td>
<td>4.516</td>
</tr>
</tbody>
</table>

1.1 Regulations in the MDS

Regulations in the MDS are represented using the CEN MetaLex ontology, an automatically generated ontology of the Basiswettenbestand (BWB, the database underlying wetten.nl), the OPMV provenance vocabulary, the W3C Time Ontology, and the Simple Event Model (SEM). Of most importance to us here is the CEN MetaLex ontology. It provides vocabulary for distinguishing levels of description of regulations along the FRBR levels of work, expression and manifestation. In the MDS, every regulation is described both at the work level (e.g. ‘the Income Tax Law’) and at the expression level (e.g. ‘the Income Tax Law of January 1st, 2013’). Every expression level resource is linked to its work via a metalex:realizes property.

Expression-level citations allow analysis of the citation network of regulations through time. Unfortunately, the targets of citations from regulations are not explicitly linked to a specific version, but only to the work identifier. This means that the resulting citation graph would be hugely disconnected: many nodes (the expressions) have only outgoing links, while other nodes (the works) have only incoming links. Consequently, measures such as betweenness centrality, clustering coefficient, connected components and network diameter will give very little information of the connections between regulations at the work-level. For this reason, the analysis presented in this paper only takes into account citations aggregated to the work level.

1. A citation from an expression level to a work, will be represented as a citation between works.
2. Two citations to separate expressions of a single work will only be counted once.
3. The highest level of detail of a citation is the article level, i.e. the most specific, uniquely and independently citable part of a regulation.
**Table 2.** Top-10 Betweenness Centrality

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Algemene wet bestuursrecht</td>
<td>7007741</td>
</tr>
<tr>
<td>2</td>
<td>Wet milieubeheer</td>
<td>2172441</td>
</tr>
<tr>
<td>3</td>
<td>Besluit omgevingsrecht</td>
<td>1667495</td>
</tr>
<tr>
<td>4</td>
<td>Besluit algemene regels voor inrichtingen milieubeheer</td>
<td>948497</td>
</tr>
<tr>
<td>5</td>
<td>Wet op de economische delicten</td>
<td>770968</td>
</tr>
<tr>
<td>6</td>
<td>Wetboek van Burgerlijke Rechtsvordering</td>
<td>696456</td>
</tr>
<tr>
<td>7</td>
<td>Wet op het hoger onderwijs en wetenschappelijk onderzoek</td>
<td>687873</td>
</tr>
<tr>
<td>8</td>
<td>Wet op het financieel toezicht</td>
<td>664934</td>
</tr>
<tr>
<td>9</td>
<td>Algemene douanewet</td>
<td>616671</td>
</tr>
<tr>
<td>10</td>
<td>Circulaire bodemsanering 2009</td>
<td>561465</td>
</tr>
</tbody>
</table>

**Table 3.** Top-10 PageRank

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Algemene wet bestuursrecht</td>
<td>0.0152</td>
</tr>
<tr>
<td>2</td>
<td>Wetboek van Burgerlijke Rechtsvordering</td>
<td>0.0117</td>
</tr>
<tr>
<td>3</td>
<td>Wet gemeenschappelijke regelingen</td>
<td>0.00803</td>
</tr>
<tr>
<td>4</td>
<td>Wet openbaarheid van bestuur</td>
<td>0.00785</td>
</tr>
<tr>
<td>5</td>
<td>Wetboek van Strafvordering</td>
<td>0.00723</td>
</tr>
<tr>
<td>6</td>
<td>Grondwet</td>
<td>0.00712</td>
</tr>
<tr>
<td>7</td>
<td>Algemene termijnwet</td>
<td>0.00668</td>
</tr>
<tr>
<td>8</td>
<td>Wet structuur uitvoeringsorganisatie werk en inkomen</td>
<td>0.00638</td>
</tr>
<tr>
<td>9</td>
<td>Kaderwet zelfstandige bestuursorganen</td>
<td>0.00623</td>
</tr>
<tr>
<td>10</td>
<td>Vreemdelingenwet 2000</td>
<td>0.00597</td>
</tr>
</tbody>
</table>

**Table 4.** Top-10 Indegree

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Algemene wet bestuursrecht</td>
<td>426</td>
</tr>
<tr>
<td>2</td>
<td>Bezoldigingsbesluit Burgerlijke Rijksambtenaren 1984</td>
<td>336</td>
</tr>
<tr>
<td>3</td>
<td>Archiefwet 1995</td>
<td>278</td>
</tr>
<tr>
<td>4</td>
<td>Werkloosheidswet</td>
<td>265</td>
</tr>
<tr>
<td>5</td>
<td>Wet op de arbeidsongeschiktheidsverzekering</td>
<td>236</td>
</tr>
<tr>
<td>6</td>
<td>Ziektekosten</td>
<td>220</td>
</tr>
<tr>
<td>7</td>
<td>Warenwet</td>
<td>210</td>
</tr>
<tr>
<td>8</td>
<td>Algemene Wet Bijzondere Ziektekosten</td>
<td>207</td>
</tr>
<tr>
<td>9</td>
<td>Wet op het voortgezet onderwijs</td>
<td>204</td>
</tr>
<tr>
<td>10</td>
<td>Zorgverzekeringswet</td>
<td>119</td>
</tr>
</tbody>
</table>
2 The Level of Regulations

In order to answer the question “What is the most important or influential regulation in the Netherlands?” we can analyse the network of co-citation between regulations as found in the MDS. Since all elements of a regulation (read articles, chapters, paragraphs etc.) are represented in RDF as part of a named graph, we can build this network by running a SPARQL query that simply returns the graph URIs of the source and target of every citation, respectively ?s and ?t:

5 Note that this query will not return all citations, but only one per source/target pair. Also, the endpoint at http://doc.metalex.eu is limited for performance reasons, so it may not return the same results as used for this analysis.


PREFIX metalex: <http://www.metalex.eu/schema/1.0#>
PREFIX bwb: <http://doc.metalex.eu/bwb/ontology/>
PREFIX dcterms: <http://purl.org/dc/terms/>

SELECT DISTINCT ?s ?s_title ?t ?t_title WHERE {
  GRAPH ?s {
    ?s_ref metalex:cites ?t_id .
  }
  GRAPH ?t {
    ?t_id a ?type .
  }
  OPTIONAL {?s dcterms:title ?s_title }.
  OPTIONAL {?t dcterms:title ?t_title }.
}

For readability purposes, we also retrieve the titles of both regulations (?s_title and ?t_title). The result is stored in two separate CSV files, one for the edges (with columns “Source” and, “Target”), and one for the nodes (“Id”, “Label”), and loaded as a graph in Gephi. Table 1 shows the network properties of the resulting citation graph. Tables 2, 3 and 4 show the top ranking nodes (documents) for Betweenness, PageRank and Indegree, respectively. The resulting graph is depicted in Figure 1, where nodes and edges are colored according to the applicable module, and size of nodes corresponds to the PageRank score.

Betweenness centrality measures the relative number of shortest paths that run through a node. The intuition is that nodes with a high betweenness centrality are important for connecting separate parts of a graph. In other words, documents with a high betweenness centrality connect different, otherwise unconnected parts of the Dutch regulations. The “Algemene Wet Bestuursrecht” (AWB), the general administrative law is high in the list as it is large, and touches upon virtually all regulations that concern the Dutch government. Environmental (‘milieu’ and ‘omgeving’), economic, and civil (‘burgerlijk’) laws have similar qualities. The ‘circulaire’ is a type of regulation that has the specific function of bringing together aspects of multiple regulations for a specific target audience.
Fig. 1. Citation network between regulations at work level.

Indegree measures the number of incoming edges to a node. This is an absolute measure of importance. Again the AWB tops the list, but other regulations reflect on penal procedures ('strafvordering'), unemployment ('werkloosheid'), healthcare ('ziekte'), education ('onderwijs') and salaries ('bezoldiging') of civil servants at the national level. PageRank [4], one of the algorithms used by Google, is a relative measure of importance. Again the AWB tops the list, but the constitution ('Grondwet') makes a first appearance, as well as regulations concerning immigrants ('vreemdelingen'), freedom of information ('openbaarheid').

2.1 Evaluation

We compared the various network measures to a list of “important” regulations, i.e. those listed on Wikipedia as belonging to the category of Dutch law. This is

\[\text{See http://nl.wikipedia.org/wiki/Categorie:Nederlandse_wet.}\]
Table 5. Comparison to laws listed on Wikipedia

<table>
<thead>
<tr>
<th>Measure</th>
<th>Recall (0.5x)</th>
<th>Recall (1x)</th>
<th>Recall (2x)</th>
<th>Precision (2x)</th>
<th>F-Score (2x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PageRank</td>
<td>0.311</td>
<td>0.444</td>
<td>0.617</td>
<td>0.309</td>
<td>0.411</td>
</tr>
<tr>
<td>Indegree</td>
<td>0.296</td>
<td>0.474</td>
<td>0.612</td>
<td>0.306</td>
<td>0.408</td>
</tr>
<tr>
<td>Degree</td>
<td>0.260</td>
<td>0.423</td>
<td>0.551</td>
<td>0.276</td>
<td>0.367</td>
</tr>
<tr>
<td>Betweenness</td>
<td>0.240</td>
<td>0.388</td>
<td>0.536</td>
<td>0.268</td>
<td>0.357</td>
</tr>
</tbody>
</table>

a very imprecise measure, but it indicates at least a basic notion of importance. We normalized all names to lower case, and removed those laws from the target set Wikipedia that do not have a direct match with any regulation in our set. The reason is that many of the regulations listed on Wikipedia are listed by citation title, rather than the full title. This reduced the list from 315 regulations initially, to 196 regulations (this is more than the 180 we could obtain by querying the wetten.nl portal directly). It should be noted that it is relatively straightforward to improve this number, e.g. by retrieving citation titles from the MetaLex Document Server, and using a simple edit distance or bag of words comparison between titles.

Table 5 shows recall and precision for PageRank, betweenness centrality, degree and in degree as they apply for varying sizes of the result set. Precision only applies in cases where the result set is larger than the target set (i.e. the length of the list of regulations from Wikipedia).

The results for this comparison shows that PageRank and in degree compete for the first place with respect to the ability to predict occurrence of a regulation in the Wikipedia category. However, PageRank performs consistently better over multiple result set sizes. Only in the case where the result set size matches the target set size exactly (Recall 1x), the in degree measure results in higher recall. It should be noted that only with a result set of 61 times the target set size, recall for PageRank and in degree reaches 100%: we need to consider approximately 80 percent of all regulations in our graph. For degree this point lies at around 50 times the target set size.

3 The Level of Articles

If we consider citations to- and from the article level, i.e. we look for an answer of the question “What is the most important or influential article in the Netherlands”, we design a SPARQL query that does not aggregate to the graph level, but considers the citing article itself:

```sparql
PREFIX metalex: <http://www.metalex.eu/schema/1.0#>
PREFIX bwb: <http://doc.metalex.eu/bwb/ontology/>
PREFIX dcterms: <http://purl.org/dc/terms/>

SELECT DISTINCT ?s_ref ?s_title ?t_id ?t_title WHERE {
```
**Table 6. Top-10 Betweenness Centrality**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Wet op de omzetbelasting 1968, Bijlage I</td>
<td>829.5</td>
</tr>
<tr>
<td>2</td>
<td>Wijzigingswet Wet luchtvaart (Regelgeving burgerluchthavens en militaire luchthavens), Artikel X</td>
<td>504</td>
</tr>
<tr>
<td>3</td>
<td>Warenwet, Artikel 1</td>
<td>492.5</td>
</tr>
<tr>
<td>4</td>
<td>Warenwet, Artikel 3</td>
<td>436.5</td>
</tr>
<tr>
<td>5</td>
<td>Wet vergoedingen adviescolleges en commissies, Artikel 2</td>
<td>423</td>
</tr>
<tr>
<td>6</td>
<td>Pensioenwet BES, Artikel 1</td>
<td>373</td>
</tr>
<tr>
<td>7</td>
<td>Administratiebesluit Bijzondere Ziektekostenverzekering, Artikel 1</td>
<td>362</td>
</tr>
<tr>
<td>8</td>
<td>Besluit inbeslaggenomen voorwerpen, Artikel 1</td>
<td>319</td>
</tr>
<tr>
<td>9</td>
<td>Rijkswet wijziging Statuut in verband met de opheffing van de Nederlandse Antillen, Artikel I</td>
<td>306</td>
</tr>
<tr>
<td>10</td>
<td>Wet openbaarmaking uit publieke middelen gefinancierde topinkomens, Artikel 2</td>
<td>294</td>
</tr>
</tbody>
</table>

**Table 7. Top-10 PageRank**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Algemene wet bestuursrecht</td>
<td>0.00262</td>
</tr>
<tr>
<td>2</td>
<td>Archiefwet 1995</td>
<td>0.00242</td>
</tr>
<tr>
<td>3</td>
<td>Wet op het financieel toezicht</td>
<td>0.00196</td>
</tr>
<tr>
<td>4</td>
<td>Zorgverzekeringswet</td>
<td>0.00175</td>
</tr>
<tr>
<td>5</td>
<td>Algemene Wet Bijzondere Ziektekosten</td>
<td>0.00167</td>
</tr>
<tr>
<td>6</td>
<td>Bezoldigingsbesluit Burgerlijke Rijksambtenaren 1984, Bijlage B</td>
<td>0.00162</td>
</tr>
<tr>
<td>7</td>
<td>Wet op het voortgezet onderwijs</td>
<td>0.00159</td>
</tr>
<tr>
<td>8</td>
<td>Wet bescherming persoonsgegevens</td>
<td>0.00150</td>
</tr>
<tr>
<td>9</td>
<td>Wet op de omzetbelasting 1968</td>
<td>0.00148</td>
</tr>
<tr>
<td>10</td>
<td>Werkloosheidswet</td>
<td>0.00147</td>
</tr>
</tbody>
</table>

**Table 8. Top-10 Indegree**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Algemene wet bestuursrecht</td>
<td>558</td>
</tr>
<tr>
<td>2</td>
<td>Werkloosheidswet</td>
<td>453</td>
</tr>
<tr>
<td>3</td>
<td>Wet op de arbeidsongeschiktheidsverzekering</td>
<td>453</td>
</tr>
<tr>
<td>4</td>
<td>Ziektekwet</td>
<td>493</td>
</tr>
<tr>
<td>5</td>
<td>Archiefwet 1995</td>
<td>398</td>
</tr>
<tr>
<td>6</td>
<td>Wet op het voortgezet onderwijs</td>
<td>364</td>
</tr>
<tr>
<td>7</td>
<td>Wet op het financieel toezicht</td>
<td>361</td>
</tr>
<tr>
<td>8</td>
<td>Algemene Wet Bijzondere Ziektekosten</td>
<td>342</td>
</tr>
<tr>
<td>9</td>
<td>Wet werk in inkomen naar arbeidsvermogen</td>
<td>327</td>
</tr>
<tr>
<td>10</td>
<td>Zorgverzekeringswet</td>
<td>326</td>
</tr>
</tbody>
</table>
As one can see, the query is very similar to the one we use for the document level citations, but instead of \( ?s \) and \( ?t \), we look for \( ?s\_ref \), the identifier of the CEN MetaLex element that cites, and \( ?t\_id \), the identifier of the cited resource. Citations in CEN MetaLex are represented inline, that is, the RDF representation does not contain explicit metalex:cites predicates on e.g. articles or members. The citations originate from resources at a lower level in the metalex:partOf hierarchy. Unfortunately the triple store of MDS (4Store) does not support SPARQL 1.1 property paths\(^8\), and ascending the metalex:partOf hierarchy via unions in the SPARQL query is very expensive (read: slow).

We therefore reconstruct the article identifier from the citation-level identifier by parsing the transparent URI of the citing element. For instance, the URI:

http://doc.metalex.eu/id/BWBR0002634/hoofdstuk/XII/artikel/33/lid/2/al/2/extref/1/nl/2012-01-01

is used to construct:

http://doc.metalex.eu/id/BWBR0002634/hoofdstuk/XII/artikel/33/nl/2012-01-01

The next step is to reconstruct the work-level identifier for the article, by removing any language tag or timestamp information:

http://doc.metalex.eu/id/BWBR0002634/hoofdstuk/XII/artikel/33

Note that we could also have retrieved the work-level identifier directly through the SPARQL query, if we had queried along the metalex:realizes predicate. However, this would have introduced yet another expensive join in the query. Table 1 shows details of the resulting citation graph, and Figure 2 depicts a rendering of the graph where nodes are sized according to PageRank, and colored according to module.

Table 10 shows the values for *in degree* of the Algemene wet bestuursrecht (AWB, administrative law) per chapter. The law itself is cited a total of 558 times, where individual parts of the law are together cited 207 times:\(^9\) 37% of all citations are to *parts* of the law.

---

\(^8\) See http://www.w3.org/TR/sparql11-property-paths/

\(^9\) Note that the document level statistics for indegree, show an aggregated number. It only counts multiple citations between two laws once, where the article level statistics count every citation.
Table 9. Power law

<table>
<thead>
<tr>
<th>Measure</th>
<th>Article</th>
</tr>
</thead>
<tbody>
<tr>
<td>PageRank</td>
<td>0.6996583</td>
</tr>
<tr>
<td>Degree</td>
<td>2.896214</td>
</tr>
<tr>
<td>Indegree</td>
<td>2.19982</td>
</tr>
<tr>
<td>Betweenness</td>
<td>3.658579</td>
</tr>
</tbody>
</table>

Table 10. Indegree per part of the Algemene wet bestuursrecht (AWB)

<table>
<thead>
<tr>
<th>Part</th>
<th>Indegree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algemene wet bestuursrecht</td>
<td>558</td>
</tr>
<tr>
<td>Chapter 9</td>
<td>52</td>
</tr>
<tr>
<td>Chapter 6</td>
<td>48</td>
</tr>
<tr>
<td>Chapter 8</td>
<td>37</td>
</tr>
<tr>
<td>Chapter 7</td>
<td>33</td>
</tr>
<tr>
<td>Chapter 10</td>
<td>9</td>
</tr>
<tr>
<td>Chapter 5</td>
<td>9</td>
</tr>
<tr>
<td>Appendix 2</td>
<td>5</td>
</tr>
<tr>
<td>Chapter 3</td>
<td>4</td>
</tr>
<tr>
<td>Chapter 4</td>
<td>4</td>
</tr>
<tr>
<td>Chapter 2</td>
<td>3</td>
</tr>
<tr>
<td>Appendix 3</td>
<td>3</td>
</tr>
</tbody>
</table>

4 Is the Law like the Web?

Both the law and the Web are man-made networks of interlinked documents. It is a valid question to ask whether the graph properties of both networks resemble each other. Or, put in another way, is the distribution of information within the body of Dutch regulations specific for the domain of law, or is it similar to the more organically grown body of information that resulted in the anarchy of the Web?

To begin to answer this question, we can consider two important properties of the Web: it is scale free (an ultra small world), and it contains a single giant strongly connected component (SCC) that contains roughly a third of all pages [1]. Scale free networks have a degree distribution that follows a power law. Table 9 lists the result of fitting various distributions to a power law function, using the igraph package in R. This suggests that indeed the degree distribution follows a power law, where $\alpha = 2.19982$, and the citation graph of Dutch legislation is scale free.

The structure of the Web resembles a bowtie (Figure 3, [1]) with at its heart a giant SCC, and incoming and outgoing nodes on the left and right, respectively. There are several smaller components that are wholly unconnected to the giant SCC, as well as tubes, that bypass the SCC, and tendrils that originate from the

---

Fig. 2. Citation network between articles at work level.

incoming and outgoing nodes. The giant SCC covers approximately one quarter of all Web pages.

The document-level network of Dutch regulations contains 14019 SCCs, one of which is 816 nodes in size, where 74 others have between 2 and 6 nodes: the vast majority are single-node SCCs. This means that although we do have a similar situation with a giant SCC, the network as a whole is not as connected as the Web. For the article-level network, the situation is even more different: 63303 SCCs with no giant SCC (maximum of 12 nodes) and 501 SCCs of size larger than one. This can be explained by the decreased likelihood of ‘random’ edges between nodes in a curated network in general, and the relatively smaller chance of an edge between articles than when edges are aggregated to document level. Indeed, Table 1 shows a much lower average degree for the article-level network. The average path length and network diameter are much smaller for
the article-level network as well, this is likely due to the much larger number of components.

5 Discussion

In the preceding we presented preliminary results of a network analysis of Dutch regulations stored in the MetaLex Document Server [2]. The networks were constructed using straightforward SPARQL queries against the MDS endpoint, requiring only minimal transformation to analyzable form. The analysis itself was performed using a variety of off-the shelf tools, primarily R and Gephi.\textsuperscript{11}

Because of the preliminary nature of this experiment, it is hard to draw any conclusions with respect to what the analysis tells us about Law. We compared the selection of regulations based on various network metrics to a sample list of regulations from the Dutch Wikipedia page. This is a relatively arbitrary selection, and the results are correspondingly in-definitive. However it does point in an interesting direction: do network metrics on citations between regulations tell us anything about the \textit{importance} or \textit{role} of those regulations? Also, it

\textsuperscript{11} For R, see http://www.r-project.org.
would be interesting to see whether citations to articles indicate e.g. a high representation of definitions in the cited article [5].

Secondly, we compared network properties of the document- and article level networks to that of the Web, and concluded that both networks are scale free, but the connectedness of regulations is much lower than that of the Web. The lower connectedness makes it easier to distinguish modules in the set of regulations. It would be very interesting to see how the results of generic module recognition algorithms correspond to actual topics in legislation.

Acknowledgments

This publication was supported by the Dutch national program COMMIT.

References


A Text Analysis Framework for Automatic Semantic Knowledge Representation of Legal Documents

Ákos SZÖKE a,1, Krisztián MÁCSÁR b and György STRAUSSZ a

a Dept. of Measurement and Information Systems
Budapest University of Technology and Economics, Budapest, Hungary.
b Multilogic Ltd., Budapest, Hungary.

Abstract. An increasing number of legal documents are published on the internet. However finding the relevant information is becoming increasingly difficult. Knowledge bases play an important role in enhancing the legal knowledge and search. In this paper we propose an automatic knowledge extraction framework that converts legal document contents into a rich multi-domain semantic knowledge base. To evaluate the approach, we implemented a software prototype and carried out a case study on the most popular Hungarian regulations.

Keywords. legal knowledge representation, semantic web, network analysis

1. Introduction

Legal professionals use legal documents as primary source of information to a large extent. An increasing number of legal documents are published on the internet that forms a greater information collection for them. However finding the relevant information is becoming increasingly difficult: the available search tools hardly keep up with the number of information sources. Information retrieval can provide a useful tool for legal information systems by serving relevant information from an information collection. Knowledge bases play an important role in enhancing the intelligence of legal knowledge and search as well as in supporting information integration. Their usage allows for more refined retrieval models in which some reasoning functionality can be applied.

In this paper we propose an automatic knowledge extraction framework that converts legal document contents into a rich multi-domain knowledge base which is compliant with the Functional Requirements for Bibliographic Records (FRBR) framework [9]. The resulted knowledge base currently describes 256 regulations (the most popular regulations in Hungary) in 3502 different time versions. To evaluate the proposed approach we carried out a case study on Hungarian legislative documents. We believe that this sample proves the applicability of the approach.

1The work is supported in part by the KMOP-2009-1.1.1. grant.
1.1. Related Work

Information retrieval is the activity of obtaining information resources from a collection that is relevant to an information need. Searches can be based on metadata, on full-text or other content-based indexing. An overview of the state of the art of innovative techniques for legal text retrieval is presented in [8]. In [3] a research is conducted on clustering of actual law firm data. An important technique of knowledge management systems is text data mining, which refers to the process of deriving high quality information from text. Text data mining can extract information from text corpora, which is a useful tool for analyzing legal documents in automated ways such as text categorization, clustering, document summarization and concept extraction. In [2] a document classification, clustering and search methodology based on neural network technology is presented. An automated text analysis of international law is detailed in [10] with four case studies. In [11] social network analysis was used to explore the cross-citations between legal cases and established some other variables to establish legal relevance between judicial decisions.

Our approach is different from the beforementioned researches. Our goal is to build semantic knowledge bases from natural language legal regulations according to the FRBR framework using text analysis techniques. Additionally, during model building several semantic vocabularies are applied such as MetaLex and LKIF (see Sec. 1.2).

1.2. Background

Semantic vocabularies are used to characterize terms, relationships and constraints on using those terms. The role of these vocabularies is to help organizing knowledge and help data integration by clarifying the terms used in the different documents and by helping to discover new relationships between terms. Some applications choose to use simple vocabularies to make the identification of the terms only. Others apply more complex ontologies with axiomatic definitions and complex reasoning possibilities. The applied level depends on the required services. In our proposed approach we use two legal ontologies, namely the metalex-owl and the lkif ontologies besides some more general ones. Additionally, we constructed a Hungarian regulation specific ontology (i.e. metalex-ext.owl) that complements the MetaLex standard. It contains 107 specific OWL entities that describes – for example – Hungarian regulation types (e.g. act, decree) as classes (5), regulation authors (typically Ministries) as individuals (78), and identification related datatypes (4). In the following we introduce the MetaLex and the LKIF ontologies.

MetaLex [1] is a jurisdiction independent XML standard for representing and interchanging the structure of legal resources. MetaLex distinguishes identification of documents on work, expression, manifestation and item level in accordance with the Functional Requirements for Bibliographic Records (FRBR) model [9]. From our point of

---

4Dublin Core, http://purl.org/dc/elements/1.1/
5Dublin Core Terms, http://purl.org/dc/terms/
7W3C Time ontology, http://www.w3.org/2006/time/
8Friend of a Friend, http://xmlns.com/foaf/0.1/
view the work and the expression levels are specifically important. The Bibliographic work is a bibliographic object created by one or more persons in a single creative process that results in a publication. The work is recognized through individual expressions. The Bibliographic expression is a realization of one bibliographic work in the form of words, sentences, paragraphs, etc. Any modification in the content gives rise to a new expression. MetaLex includes an OWL schema that specifies concepts such as BibliographicWork, BibliographicExpression in order to make explicit to which level a metadata pertains. Besides the vertical relationships the MetaLex OWL schema includes a number of event type definitions on linking metadata to events (such as creation, enactment, repeal) to make this horizontal relations explicit. The identifiers of documents and its parts or metadata statements about the documents are expressed as Internationalized Resource Identifiers as defined by RFC 3987 [4].

A legal ontology plays an important role to articulate knowledge in legal knowledge systems. The Legal Knowledge Interchange Format (LKIF) [6] is proposed to provide a knowledge representation formalism that is a part of a larger architecture for developing legal knowledge system. LKIF is developed considering other ontologies and frameworks, such as Core Legal Ontology (CLO) [5] and Language for Legal Discourse (LLD) [7], and it is intended as a core ontology for law. LKIF also comply with the current Semantic Web standards to enable legal information serving via the web.

2. The Emerald Knowledge Extraction Framework

The Hungarian legislative documents consist of free text (HTML), but also contain various types of structured information in the form of metadata. Our aim is to extract the information both the metadata and the free text and turn them into a rich knowledge base.

![Figure 1. Architecture of the Extraction Framework.](image-url)
2.1. Architecture of the Extraction Framework

Figure 1 gives an overview of the knowledge extraction framework. Its main components are: SourceManager which is an abstraction of local or remote legal document sources, ExtractorManager which extracts the information from the documents, and KnowledgeBaseManager which builds a knowledge base from the extracted data according to the predefined document metamodel. All of these managers are different stations of the processing workflow.

Since every source has different characteristics – such as identification method, versioning method, citation styles, and metadata – every source (Source\textsubscript{1}...Source\textsubscript{n}) requires different abstract data model. These abstract models define which kind of data can be extracted from the document with different extraction rules and these are used to build up the concrete data models (i.e. semantic knowledge models). Due to the similarity of the documents, abstract models are constructed using inheritance. It enables extending existing abstract models easily.

The core of the framework is the ExtractionManager which manages the process of passing legal documents from the SourceManager to the extractors and delivers their output to the KnowledgeManager. The Extractors are different extraction jobs. They identify and take the specific type of information out from the text. Naturally, the different type of documents (e.g. act or decree) often requires different extraction rules. The Parsers support the extractors by determining datatypes, geographical data, and language. The framework currently consists of four extractors:

- **ConceptExtractor**: all documents are analysed and annotated with the EuroVoc vocabulary\textsuperscript{10} which is the EU’s multilingual, multidisciplinary thesaurus covering the activities of the EU. It contains terms in 22 EU languages. For analysis we applied the magyarlanc NLP (Natural Language Processing) tool\textsuperscript{11}. Because of the nature of the automatic processing, manual revisions may necessary due to the positive / negative false results.
- **MetaDataExtractor**: all documents are analyzed to extract information according to the pre-defined data model. This extracted information includes such as metalex-owl:Author, dcterms:title.
- **CitationExtractor**: documents contain citations to other documents that we represent with the metalex-owl:cites property.
- **VersionExtractor**: every work level and expression level documents are characterized with different event types such as metalex-owl:LegislativeModification which describes the circumstances of the legislative change. The work and the expression level documents are connected with a metalex-owl:realizes property.

The KnowledgeBaseManager builds a knowledge base (i.e. RDF triples\textsuperscript{12}) from the extracted data with the predefined mappings in the following manner. First the corresponding URI of the processed document is created from some document characteristics such the local identifications such as work category (e.g. act), issueDate and issueID.

\textsuperscript{10}EuroVoc, http://eurovoc.europa.eu/
\textsuperscript{11}Homepage of magyarlanc, http://www.inf.u-szeged.hu/rgai/nlp
\textsuperscript{12}Resource Description Framework, http://www.w3.org/TR/rdf-primer/
Then, the predicate URI is selected from the previously mentioned vocabularies (e.g. metalex-owl, dcterms). Third, the objects of the triples are built up from the extracted data to generate suitable URI references or literal values. Finally, these triples are uploaded to the knowledge base store (i.e. triple store). The results of the process can be browsed and analysed with RDF Browsers, SPARQL\textsuperscript{13} clients as well as a custom client application (Emerald WebVisualiser\textsuperscript{14}).

2.2. The Emerald Knowledge Base

In the following an excerpt of the constructed data model of a regulation is presented. The presented data model is made up of RDF statements of the form of \((subject, predicate, object)\). The \textit{subject} is the resource described by the statement. The \textit{predicate} is the property used to relate the subject to the object. The \textit{object} is the value of the property as it holds for the subject.

The code snippet in Code 1 shows the 5 main building blocks of regulation data model. The square brackets represent the generated subject URIs of the blocks. The predicates and the objects of the statements are indented. The prefixes (e.g. dcterms) point out the semantic vocabulary from which a given predicate or object is derived.

\begin{verbatim}
[WorkURI]
 rdf:type metalex-owl:BibliographicWork
 metalex-owl:resultOf [WorkCreationEventURI]
 rdf:type metalex-owl-ext#:act
 dcterms:title "<Title of the legal document>"^^<xsd:string>
 metalex-owl:Author [WorkAuthorURI] # e.g. some ministry
 metalex-owl:countryCode "<jurisdiction-code>"^^<xsd:string> -- e.g. "hu"
 metalex-owl:issueID "<document-id>"^^<xs:integer> -- e.g. "11"
 metalex-owl-ext:issueDate "<date>"^^<xs:integer> -- e.g. "1960"

[ExpressionURI]
 rdf:type metalex-owl:BibliographicExpression
 metalex-owl:resultOf [ExpressionCreationEventURI]
 time-modification:in_force [ExpressionTimeModificationURI]
 metalex-owl:languageCode "<language code>"^^<xsd:language> -- e.g. "hun"
 metalex-owl#realizes [WorkURI]
 metalex-owl:cites [WorkURI]

[WorkCreationEventURI]
 rdf:type metalex-owl:LegislativeDelivery
 metalex-owl:participant [ParticipantURI] -- e.g. "Parliament"

[ExpressionCreationEventURI]
 rdf:type metalex-owl:LegislativeModification
 metalex-owl:participant [ParticipantURI] -- e.g. "Parliament"
\end{verbatim}

\textbf{Code 1}: Data model excerpt of regulations.

The first and the second blocks represent the FRBR work and expression level entities (see \texttt{rdf:type} properties) of a regulation respectively. The work contains the title,
the category of a work (e.g. act), the issue id (issueID) and date (issueDate) and
the author of a regulation. In Hungary a triple of category of work, issue date and issue
id identifies a regulation. The expression involves reference to the work (realizes),
citations to other regulations, language of the expression and reference to the temporal
description of the resource (in_force) validity that defines the temporal version of
the regulation. Since every bibliographically relevant change emanates from legislative
events these events are represented in the third and the fourth blocks and these are refer-
cenced from the work and the expression (resultOf). The participant of the event is
described in the adequate property.

3. Experiments

A case study was carried out to evaluate our method on real-life situations. The case
study was performed using the previously detailed approach and our prototype (named
EmeraldRdfExtractor).

3.1. Context and methodology

In the case study we primary focused on citation analysis as the concept extraction anal-
ysis can easily fill a paper itself. Building citation networks not only can improve search
results, but they have other significance also. Citation analysis is a way to determine
the 'importance of regulations’ i.e. which are the most central regulations of the law
system. There are several metrics to measure importance such as citation counts, HITS
or PageRank (see [11]). These measures can be also combined with queries to provide
better results. Citation analysis also provides guidance in change analysis of the law: if
we change (a part of) a regulation then which other regulations are affected and should
be amended. Although legal relevance correlates with citations, some studies points out
the importance of different publications in legal magazines and other sources such as
references in commentaries.

In the case study, first, we selected the most popular 256 legislative documents (i.e.
work level document) in Hungary which contained 3502 versions (i.e. expressions). All
of these documents are represented in HTML format. Then, we created 6 document data
models that reflect the different type of legal regulations (act, decree, etc.). Third,
we identified and defined 67 extraction rules in the form of regular expressions includ-
ing 32 kind of citations. The extractor currently supports document level and document
part level extraction but it only works on singular citations (i.e. citation intervals are
excluded).

3.2. Data collection and Results

The 3502 documents resulted in about 524 thousands triples due to the extraction. An
expression contained about 150 triples on average. About 73 percent of the triples de-
scribed citations (385 thousands). The 256 regulations (work) cited 1672 documents with
more than 7350 links (multiple citations not considered). Next, we only summarize some
highlights of the citation network analysis in Table 1.
Table 1. Statistics of the citation analysis

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Definition</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-degree</td>
<td>the number citations from a document</td>
<td>[1, 129]</td>
</tr>
<tr>
<td>Out-degree</td>
<td>the number citations to a document</td>
<td>[1, 121]</td>
</tr>
<tr>
<td>Average degree</td>
<td>the number citations from/to a document</td>
<td>4.4</td>
</tr>
<tr>
<td>Network diameter</td>
<td>the maximum citation-distance between all pairs of documents</td>
<td>9</td>
</tr>
<tr>
<td>Average path length</td>
<td>the average number steps along the shortest path for all documents</td>
<td>3.4</td>
</tr>
<tr>
<td>Modularity</td>
<td>the strength of division of the network into communities</td>
<td>0.39</td>
</tr>
<tr>
<td>Number of communities</td>
<td>detected communities using $resolution = 1.0$</td>
<td>15</td>
</tr>
</tbody>
</table>

From the analysis, we can say that the most influential regulation considering citations (the highest indegree: 129) was an act on the restructuring of the government organisation (act, 2006/CIX). The greatest number of citations (the highest outdegree: 121) was an act about financial institutions (act 1996/112). The framework can process a documents per 0.5 second on a 2.2 GHz dual-core machine with 4 GB memory. It includes reading the document, processing and uploading the triples into the triple store.

Figure 2 shows a filtered view of the citation network (nodes with less than 35 degree are filtered out) to provide a visual overview of the citation network. In this figure the size of the nodes correlate with degree of the nodes and the different communities (after filtering 13 communities are remained) have different colors. This figure also points out the relative weights (the given community per all communities in terms of degree).

Figure 2. Filtered citation network (left) and its communities with their relative weights (right).
4. Conclusions

To improve the search accuracy on legal corpus, we proposed a text analysis framework that builds semantic knowledge representation from natural language legal documents. The knowledge representation not only makes the accuracy of (semantic) search possible but enables the formal articulation of domain knowledge at a high level of expressiveness that enables to specify more detailed queries. To obtain a proof-of-concept we implemented a prototype, named EmeraldRdfExtractor, and carried out a case study on Hungarian regulations. The presented method is a combination of the present theories and methods thus it leads us to generalize our findings beyond the study.

References

Graph-Based Linking and Visualization for Legislation Documents (GLVD)

Dincer GULTEMEN, Tom van ENGERS

Abstract

Law requires the consistency of its components called sources-of-law. Therefore, the consistency verification is essential through the audit of inter-regulatory interactions, which may also be supported through the cooperation of network approach and the semantic technology applications. Because the interactions between documents representing sources-of-law may be conceptualized as source-network-of-law (SNL), graph-based approach may provide a ground to audit these interactions and to analyze a possible multi-layer semantic network with increased granularity. Therefore, this paper is aimed to construct a SNL model of interactions in a composite graph structure as well as processing the model in parallel with the development of a semantic tool (GLVD) which may constitute a basis for more advanced future applications requiring inter-regulatory interactions to be detected, to be labeled for data linking or to be visualized in a graph formation. Within the scope of the initial version, a basic type of inter-regulatory interaction, which is formed through explicit references embodied in legislation (statutory law) documents, is taken into consideration. Then, a complete data processing pipeline including appropriate “text to XML parsing” and “XML to visualization” is aimed to be provided. Therefore, firstly, GLVD contains a proper “text to XML” parser module producing separate XML documents tailored for algorithmic paragraph-article level document linking through self-labeling. Secondly, “XML to visualization” module provides a combined arbitrary cyclic-tree graph structure through a semantic XML linking. Finally, the visualization of composite graph is realized in line with the principles of conceptual graph visualization.

Key Words: Legislative Text Processing, Legislative XML, Network Analysis, Graph-based Knowledge Representation, Information Visualization.
1-Introduction

The concern for the idea of transferring semi-structured information and the use of semantic technologies has increased during recent years. It may be observed that one of the very basic aims of the Information Communication Technologies (ICT) in business process management is the reduction of the repeated steps of workflows. Previously, the use of centralized databases with concurrent applications has served properly to the aim of process reduction through the interoperability inside of the institution. This achievement has also stimulated a latter idea of inter-institutional business process reduction through interoperability. Therefore, as a single legislative authority and basic harmonizer-hub for the orchestration of inter-institutional law-making business process, it may be observed that the parliaments are usually demanding the semantic technologies with the purpose of interoperability in the first place. Secondly, the very basic legislative life-cycle for statutory law and its interactive relation to the other components of sources-of-law can also be conceptualized as complexly interlinked sources-network-of-law (SNL), and the possible use of semi-structured information with the purpose of consistency verification and legal analysis also motivates the use of semantic technologies within the parliaments.

Law requires the consistency within the SNL. Therefore, the consistency verification is essential through the audit of inter-regulatory interactions, which may also be supported through the cooperation of network approach and the semantic technology applications. Therefore, the scope of our paper includes the statutory law (legislation) level SNL which may be created through the transformation of related documents into semi-structured legal information, and semantically enriched law documents may then be suitable for semantic analysis accordingly. SNL, which is created through explicit reference linking and represented as a conceptual graph formation with node granularity at the paragraph-article level, may provide a ground for audit of interactions and may also be improved for possible multi-layer semantic network analysis with increased granularity. Besides, the possibility of case-law to be included to peripheral to the graph also provide more composite notion of SNL and network analysis.

This paper is aimed to construct a SNL model of interactions in a composite graph structure by also processing the model in parallel with the development of a semantic tool (GLVD) which may constitute a basis for more advanced future applications requiring inter-regulatory interactions to be detected, to be labeled for data linking or to be visualized in a graph formation. A complete data processing pipeline may then be provided through an appropriate “text to XML” parser module producing separate XML documents tailored for algorithmic paragraph-article level document linking through self-labeling, and “XML to visualization” module providing a combined arbitrary cyclic-tree graph structure through a semantic XML linking and the visualization of composite SNL graph from multiple documents.

2- Literature Overview

We consider that it may be concluded that scientific literature at the intersection of graph theory and law can be classified into macro-level and micro-level network analysis. Macro-level analysis is exploring the overall structure of legal citation network including cases,
statues and other legal authorities while micro-level analysis is concentrating on more granularity, logical-semantic aspects and precision in node-link generation. Fowler et al. (2006) concentrated on “per curiam” (majority decisions signed by the court) US Supreme Court Opinions (USSCO) to measure case centrality, based on a principle of “stare decisis”, in terms of required relevance-authority to serve as a precedent. Then, the network was constructed with 26681 USSCO (1791-2005) as nodes, with directed citations links, and with importance scores as a measure to define most legally relevant and how central a case to law at the court. A variant of eigenvector approach was used to calculate importance score and to rescale the priority node-rank by relating inward and outward links mathematically as two interacting eigenvectors. Smith (2007), on the other hand, concluded that the network structure of U.S case law has the highly skewed distribution of citations, which can better be described by the power law distribution, indicating that very few cases get the vast majority of citations where cited cases fall rapidly with the number of citations. This pattern explained with the simultaneous mechanisms of preferential attachment and best fit competitive attachment. Another study by Boulet et al. (2011) was to understand the dependencies between domains of French legal codes and to provide objective definition, characterization and audit for the concept of complexity, which may also cause law to be considered as unconstitutional. Then, the graph of citations, which is produced by cross-citations among “legislative drafting-procedure” based legal texts, were studied based on measures like small world indices like clustering coefficients of “local connectivity” and “global connectivity” including diameter, the mean of shortest paths, characteristic path-length and following centralization measures. The resulting conclusion was implying a “concentrated world” with lower density than “small-worlds” but a higher density than a “random graph”. Besides, “modularity partitioning” measures were concluding a hidden interdependence between clustered legal domains including the existence of rich-club. Finally, visualization on hemicycle-like plot based on the statistical reduction of “dissimilarity measures” provided through Euclidian dissimilarity between vertices has also been provided. Later, Winkel et al. (2011) tried to validate the network approach for the Dutch case law to find “most fit” case where all cases can not be perceived as equally important within growing number of published sources of law. Then, a citation network of 15053 Dutch Supreme Court (DSC) Decision (1965-2008) with 106559 inward and outward citations was parsed and domain experts were likely to agree with results. A second analysis including Dutch case law showed that DSC was far more authoritative than lower courts, low degree courts were rare to be cited and PageRank algorithm based on a trust notion was not helpful as betweenness centrality and incoming links because the importance of the referrer is less important in legal field. The results were in line with continental legal tradition where network focused mostly on procedural law related citations because Supreme Court judges primarily judicial procedures rather than material facts. Besides, the number of references reasoned as “breaching the previous line of thinking” or moving from “suspicion of quilt” to “risk liability” might be considered as indicators for legislator to fill the legal gap and it has dropped dramatically after related legislation.

4Radboud Winkels, Jelle De Ruyter and Henryk Kroese; Determining Authority of Dutch Case Law; Legal Knowledge and Information Systems, Vol. 235, pp. 103-112; 2011.
The other group of network analysis concentrates on more granular structure for legal texts formed into graphs which we call here as micro-level network analysis. The suggestion on the effective treatment of legal documents is based on the ability to relate documents of different types through the use of document graphs, where nodes represent the meaningful units of texts (textual objects) linked with edges of logical relationships, can be traced back to Bench-Capon (1997). Accordingly, legislative documents within a hierarchical structure were related based on graph modification specifications which also provide a basis for the consistency of document network. Then, a single composite structure with the possibility of linearization could be concluded. Micro-level networks can be limited horizontally in legal hierarchy of the sources (such as statutory law documents) with the purpose of legislative compliance, or vertically including various sources of law within a domain with the purpose of regulatory compliance. Therefore, with the recent study of Hamdaqa and Hamou (2011) aiming vertical regulatory compliance (which was also horizontally limited by statutory-administrative level North-American laws and regulations), an application of graph theory was suggested for the detection of duplicated, overlapping or conflicting provisions based on regulatory dependency patterns such as defined-by-defined-by, amend-amend, amend-use, generalized amend-amend. Law parser of the supporting tool of CompDSS (Compliance Decision Support System) was to parse the provisions identifying all parts of document in hierarchical tree form indexing with their location path from title as they are cited in legal documents, and to parse citations based on automata to recognize possible and regular expressions to capture standard citations forming the data model. Then, the relations were processed based on Brill’s transformation-based learning tagger to label citations as assertions (definition, specification, compliance) and more frequently amendments (by insertion, by deletion, by striking, by redesignation). Finally, the graph was specified as $G = (V, E, R)$ where edges for citations (E) relation type-labeled with (R) between provisions (V) where its data model were expressing the act with provisions having hierarchically formed provision element with leaf elements and composite elements. The graph was also visualized by the related tool module in DOT scripting language and Graphviz with “neato” approach.

With our aims partially overlapping with previous research conducted within micro-level network analysis, our approach here is to achieve a complete data processing pipeline application for micro-level network construction with a more technical focus including parsing and visualization making use of the semantic technologies. Then, it may also be expected to bridge it with macro-level network analysis with increased granularity through.

2- A Technology for Interoperability

Information transfer from one software application to another requires certain communication channels, protocols and information representation conventions. Hypertext Transfer Protocol-Hypertext Markup Language (HTTP-HTML) mechanism is a common mean for transfer. Although procedural markup language HTML has presentational advantages over descriptive markups like Extensible Markup Language (XML) or its predecessor SGML, it is not capable to fit units into complete and meaningful content structures by using desired tag names which are initiated, ended properly and nested within each other hierarchically. Being modular and

---

6Mohammad Hamdaqa, Abdelwahab Hamou-Lhadj; An approach based on citation analysis to support effective handling of regulatory compliance; Future Generation Computer Systems 27 (2011) 395-410.
7Airi Salminen, Frank Tompa; Communicating with XML; Springer Science+Business Media, 2011.
verbose by design, XML is for reflecting the tree structure of the document data with an embedded mark-up consisting tags and syntax in both human and machine readable way, leads family of technologies, XML complied HTML (XHTML) and serves as a open-standard basis for next generation semantic technologies.

Because semantic analysis usually involves in irregular, deeply hierarchical and recursive data; tree-structured information offers a better management in semantic processing than it is the case in relational information structure. This does not, on the other hand, mean that the XML technology is a rival to database technologies which can still provide advance features of transaction management, security, multiuser access, scalability, and so on. However, it should also be noted that the development of native XML databases or XML-enabled databases may also serve as a proof for the contribution of XML technology to the demand for interoperability and data migration. The rise of XML has also led its advances from Resource Description Framework (RDF) to Web Ontology Language (OWL). The semantic features built on XML structure to integrate knowledge representation or ontologies result in the main foundation of the semantic web RDF, which is also known as semantic XML. However, we may still want to use XML based applications rather than RDF to link the documents making use of the higher level universality in current standardization of XML, and therefore, the capacity for interoperability.

3- An Application for Semantics

The very core of the idea of “semantic web” lies on the concept of intelligent data rather than intelligent application. RDF is more intelligent than the XML in the sense that it contains already embedded relations, and OWL is more and more because the data also carries the intelligence of ontologies. But on the other hand, the evolution line of semantic technologies also refers to a trade-off between the statically structured RDF triples network where additional labeling on XML will be required, and flexibly structured XML network where additional applications for data linking will be required to play with XML better. Besides, some experimental relations based on predefined patterns may not be decided to be annotated and linked yet. Therefore, in this paper, the focus is more on application-level semantic processing possibilities for legislative XML with the advantages it offers in interoperability. Then, it is expected GLVD achieve “text to visualization” processing pipeline through XML production provided by a module of itself, or other proper XML documents. It should also be noted that once GLVD defines the linking, it may be extended with a module to achieve RDF/XML serialization which may also be considered as a requirement based on the graph construction limits imposed by processing power. Then, GVLD will still be promising because graph construction does not depend on a single composite document, but on a single composite graph. That is, a group of documents based on legal domains can be processed and agreed upon relations can be annotated. A further visualization with less processing power may then be maintained based on intelligent data rather than intelligent application.

---

4- Text to XML Parser

A parser is a stream reader application to convert input tokens to output tokens based on grammatical rules. It is both possible to produce a parsed document with one token containing tree-structured object or series tokens passed between cascading parsers. For example, as simple parser, a lexical analyzer emits one token per word based on simple lexical rules, and then the tokens are cascaded by another parser. The parsing of XML documents by dissecting into elements, on the other hand, is conducted by XML processors which are bridged to the application through Application Programming Interfaces (APIs) such as Simple API for XML (SAX), Document Object Model (DOM) or XML Object Model (XOM). While the access of DOM to XML document structure is based on the creation of a tree structure, the access of SAX to the document is event driven. The document structure is created by callback functions instantiated by event handlers such as starting tags. Then, parsing process is not traversed as it is the case in DOM. This kind of dual method for XML processing is apparently caused from a space (memory)-time (processing) tradeoffs. It is estimated that DOM needs three times more space than SAX requires, while the structure of the document is ready at once and available for application processing when DOM is used.

XML Object Model (XOM) confronting parsers, on the other hand, tries to combine the advantages of both SAX and DOM interfaces. The very basic design principle of XOM parsing is to provide a fast-enough and small-enough parser. The underlying SAX parser to read the document provides a fast-enough speed for convenience, while it is two or three times smaller with respect to memory footprint. Because of its hybrid advantages, XOM is decided to be the API for the GLVD module to parse text to XML which is tailored according to our legislative purposes.

Legal information capture is done by a markup through several global attributes which carry basic information on the legal aspect of the document. Tags may be created based on elements containing metadata, specific structures or ontology specific text-term mining. XML schemas and Document Type Definitions (DTDs) are also very useful features of XML defining markup rules on base, derived and extended data types, facets, value limits, enumeration, patterns and repeating element occurrence specifications for the sake of data orientation, structural organization, granularity, intended processing, standardization, reuse, performance and, finally, interoperability. With this purpose, the use of CEN Metalex standard requires simple jurisdiction-neutral vocabulary for legal XML processing, avoiding

---

16Elliotte R. Harold: XOM Design Principles; Proceedings of the Extreme Markup Languages-2004 Conference; Montréal, Quebec, Canada.
any specific legal jargon and a further mapping for element “part” to layer groupings like chapter, section, part and paragraph\textsuperscript{19}. However, for the initial version of GLVD, we hold the view of using a self-descriptive XML and non-rigorous use of descriptive element and attribute containers delaying the use of exact CEN Metalex Standard schema.

At the core of the GLVD module for XML production, the concern is to link XML elements based on self-labeling. That is, elements should have an attribute value in natural language to define their structural position within XML tree structure. For example; if a paragraph of an article contains a reference pattern like “Article X of Law number Y”, then this pattern should match the attribute label of article X which should be specified as “Article X of Law number Y”. For the initial version, GLVD supports single reference pattern for explicit references. However, within the scope of previous researches, a number of patterns have already be studied for text retrieval such as basic references, references to ranges, multiple references, zooming-in layered references and zooming-out layered references\textsuperscript{20}. It is possible to improve GLVD parser by creating reference retrieval patterns repository which will be replaced with the compliant reference matching structural attribute labels.

As an initial step for legal document to XML parsing, the creation of domain specific XML tag names are organized based on the ordinary organization of legislation document terminology in English. The assumptions governing self-descriptive markup should be as follows;

- The elements of the markup are decided to be as “Regulation”, “Law”, “Section”, “Part”, “Article”, “Title (h2)”, “Paragraph (p)” and “Regulation Date”.
- The creation of these tags should be in a nested way. However it should also be mentioned that there should also be some elements nested directly by upper wrappings like titles of articles, and section or part descriptions.
- Because the granularity of semantic linking (which will be done in second phase) should be at paragraph-article level rather than at document level, self-labeling should be done accordingly using the attributes of the paragraph elements.
- The labeling should be done in a way that the attribute values should be exactly in a way that it appears on the reference texts in article paragraphs. With this purpose several counters should be set as variable and algorithmic assignment should be achieved properly.
- For the instantiation of there tags, basically two basic event groups exist in the text file. The tags may be instantiated based on the counter of empty lines or based on some specific word-phrase patterns exist in text file. Wrapping triggers for “Law”, “Section” and “Part” tags are decided to be based on empty line processing while it is inline processing for “Article” and “Regulation Date” tags. Accordingly, two consecutive empty lines will be a trigger for “Section” elements and one empty line will be a trigger for “Part” elements. “Article” instantiation will be based on the word “Article” appearing in the beginning of the line and “Regulation Date” tag will be instantiated with the word “Date” covering the next character group after the word “Date”.


\textsuperscript{20}Emile de Maat; Making Sense of Legal Texts; SIKS Dissertation Series No. 12-26; Ridderprint BV, 2012
Those assumptions are necessarily jurisdiction dependent. However, it is still possible to improve GLVD by providing an algorithm repository based on markups of different jurisdictions. The resulting markup based on current assumptions is expected to be as it is as follows;

```xml
<?xml version="1.0" encoding="utf-8"?>
<Regulation ID="r">
  <Law ID="Legislative Document y">
    <Section ID="Section s of Law Number y">
      <Part ID="Part p of Section s of Law Number y">
        <h2 ID="Title Number t1-t2 of Law Number y">TEXT</h2>
        <p ID="Paragraph Number 0 of Article Number x Law Number y">TEXT</p>
        <Article ID="Article x of Law Number y">
          <p ID="Paragraph Number z of Article Number x of Law Number y">TEXT
            <Regulation Date>TEXT</Regulation Date>
          </p>
        </Article>
      </Part>
    </Section>
  </Law>
</Regulation>
```

The text file to be parsed, therefore, should be organized in a format to be recognized by the parser. The assumptions governing text file format should be as follows;

- There should be one space line after parts and two space lines after sections of the text.
- The lines should contain the entire “Paragraph”. Length limit for a single line is set as 500 characters.
- There should be one character space after the words specified as “Law Number:”
- There should not be characters specified as “.” or “;” in the title lines.
- The titles of articles should be located above the articles while descriptions of parts and sections should be located below the section or part titles.
- Explicit references should be compliant with “Article” element attributes.

The pseudo-code of parser algorithm for empty line processing can be described as follows;

```plaintext
While (The line is not null) {
  Do inline processing;
  If (The line is not empty) {
    Create <p>paragraph</p> or <h1>title</h1> by searching empty lines
    Add paragraphs to the Article or Part element directly based on inline processing
  }Else
    BlankLines +=1;
    Add Article element to the Part element
    Set an automatic attribute for Article element label
    Add Part element to the Section element
    Set an automatic attribute for Part element label
    Clear Part element for the next iteration
    Clear Article element for the next iteration
  }
If (blankLines==2) {
  Add Section element to the Law element;
  Set an automatic attribute for Part element label
  Clear Section element for the next iteration
}
```
Some tags may also be instantiated from inline processing to form the article or sub-article level structure and to wrap core information based on their positions of the words specified such as jargon words, dates and numbers. In our case, the `<Article>` and `<Regulation Date>` tag is decided to be triggered through inline processing. The pseudo-code of parser algorithm for inline processing and the text formation can be described as follows;

```plaintext
While (The line is not null) {
    If (The word “Law Number” is detected in any line) {
        Get the value of Law Number
        Set an attribute for Law element label by using the value of Law Number
    }
    If (The line is staring with the word “Article” is detected in any line) {
        Add Article element to the Part element
        Set an automatic attribute for Article element label
        Clear Article element for the next iteration
    }
    For each line {
        Create a new element and a string value to be replaced
        Find the position and length of jargon to be replaced in the line
        Add the replaced jargon to the element
        Add the element to the built data
        (After replacement strings have been defined, the function should be iterated)
        Iterate the parser equal to the jargon+length
        If (The jargon is “Date”) {
            Create a new element called RegulationDate to wrap all dates mentioned
            Set counter to zero to retrieve character group after Date
            While (The initial position of the jargon+length+counter+1 is not whitespace)
                Add the character to the element
                Increment the counter
            Add RegulationDate to the built data
        } (After replacement strings have been defined, the function should be iterated)
        Iterate the parser equal to the jargon+length+counter+1
    }
}
```

4.1- Experimentation of “Text to XML Parser”

The experimentation of parsing has been conducted through two sample legislation texts. The texts are tailored for this paper to be able to achieve, visualize and present the simple possible linking through 3 explicit references, 2 of which is an external and 1 of which is an internal reference. Then, an appropriate production of -still separate- XML version documents from example legislation texts has been achieved for further XML visualization.

5- XML Graph Composition and Visualization Module

It is suggested that a query mechanism by using BG homomorphism may produce more effective solutions. Therefore, in the second module of the GLVD, SNL is tried to be conceptualized for legislation and automatically created in a formation of basic conceptual graph (BG) of subsumed graphs, by this way it is also expected to contribute various level knowledge engineering applications like graph-based knowledge representation and
reasoning, multi-level legal knowledge representation, legal and legislative verification, regulatory affairs surveillance, documental profiling and so forth.

Basically, SNL is a normal BG where the composing nodes are representing different entities. The composing node sets are C (set of concept types) marked by M (set of individual markers excluding generic type marker “*”), and R (set of relation types) are used to represent facts which are an assertion to relate entities (concept nodes) ordered according their types and subtypes (type(c), marker(c)=[c:m]|ccC, mcM). It is also possible to use relation signatures (type(r)) to map a relation symbol (r) to the “maximal concept type” (the most specific inclusive type of its arguments (σ(r) ∈ Cj for 1≤j≤k) based on the arity (k number of nodes involved)) of the relation. Based on the vocabulary explained above, the BG is obtained as a graph G=(C, R, E, I) by also adding the edges (E) connecting the mentioned nodes and the labeling function (I) for the edges.

SNL forms finite, undirected, and multi-graph where two different sets of nodes may be combined by several edges. The attribute labels of entities can be organized as individual markers to reflect the hierarchical relation of entity types which in fact forms tree-structured legislation XML elements. Then, a text-value based automatic analysis to establish a relation type will be more convenient to apply. For the sake of simplicity in graph composition and visualization, the relation signature to form a reference relation is decided to be:

- r=paragraphHasReferenceOf(Paragraph, Article) with mapping
- σ(r): ([Element:Paragraph], [Element:Article])

and, the relation signature to form a hierarchical relation is decided to be:

- r=elementIsTheSubOf(Element, Element) with mapping
- σ(r): ([Element:*], [Element:])**

After the creation of domain specific XML document files,

- C is formed by the hierarchical mapping M composed of Regulation / Law / Section / Part / Article / h2 / p / Regulation Date
- M also creates an implicit siblingsOf relation (through local variable x and global variable y) based on their same level in hierarchical mapping composed of Legislation Document y / Section x / Part x / Title Number x of Law Number y / Paragraph Number x of Law Number y and Article x of Law Number y.
- R is formed as paragraphHasReferenceOf and elementIsTheSubOf relations.

The development of GLVD linking and visualization module is based on two components interacting through an interface which is designed for the sake of increased modularity or visualization and linking functions. By this way, linking algorithms can be changed easier without changing the visualization component.

Firstly, within the scope of linking component, a parent “file node” is created for XML files, and the file is parsed to get its root element to append it to the parent “file node” as a child “element node”. Then, the rule definitions to create recursive nodes in a parent-child relation will be based on the number of files in a folder creating legislation structure and the hereditary parent-child relationship of individual XML Document. Starting from the parent of

21Michel Chein Marie-Laure Mugnier; Graph-based Knowledge Representation, Computational Foundations of Conceptual Graphs; Springer-Verlag, 2009.
root element node, and each node creates its children as long as the XML tags contain wrapped tags.

Create a global root element
For all files in the folder
   Parse each file and get the local root element
   Append local root to the global element with a given child name (Attribute Value)
Create a root node for global root element
Create a model: Iterate node generation in a nested way as long as elements has children
   If the element is an element having children,
      Read XML child elements
      Create nodes for each child
      Add each created node to the global control list

Then, the relations through the processing and the comparison of the content of created nodes are established. This will inevitably require all nodes to be created and the global control list to be produced.

Create an empty relation list for explicit reference
Process and compare the content of each possible pair of nodes in global control list
   If value of a node contains the name of another node (Attribute Value)
      If this node is a “Paragraph node”
         Add both nodes to the explicit reference list

At this stage, the nodes outside the created explicit reference relation and the parental relations of the nodes involving explicit reference relation are cut off. This will require all children names data (Attribute Values) of nodes in the global control list should be processed and stored. Then, for each node, it should be checked if children name data contains the node names in explicit reference list.

For all nodes in the global control list
   Find children names of each node nested and store this data in the node content
   If children name data does not contain the nodes in the explicit reference list
      Remove the children of that node

Finally, the previously created XML graph structure and further labeling for the predefined relation types constructed by the linking component is transferred to the visualization component. The development of the visualization component is based on Java Foundation Classes (JFC) including APIs like AWT / SWING / Java2D and it is also expected to be enriched with new functions and animations in the future. Translation-transformation with translation-animation “move to center” based on user controls have currently been provided.

The resulting snapshot shown in “Figure-1” represents a view for one internal and two external links owned by “first experimental law”. A view for two external references of second experimental law is presented in “Figure-2” and an overview for our tailored SNL is presented in “Figure-3”.

6- Conclusion and Future Work

In line with our aims with GLVD, we could provide a prototype tool to produce separate legislation XML documents in a format based on self-labeling attribute generation to link combined arbitrary cyclic-tree graph structure through a semantic XML linking in run-time and to visualize the composite SNL graph structure. Therefore, very basic consistency verification through the audit of inter-regulatory interactions is supported and represented by a conceptual graph formation with node granularity at the paragraph-article level.

With the future work GLVD may be improved by:
- Experimentation on large scale real SNL graph to be conducted.
- Increased relation types to be enhanced with more semantic features through creating reference retrieval patterns repository which will be replaced with the compliant reference matching structural attribute labels.
- CEN Metalex compliance to be maintained to cope with existing dataset with a satisfactory sample size.
- Case-law to be included to peripheral to the graph providing more composite notion of SNL and network analysis to estimate the need for legislative action to be taken against the court tendencies.
- An algorithm repository based on markups of different jurisdictions to be provided.
- Multilevel network analysis to be tried to achieve with the use of graph based knowledge representation.
- More dynamic view for the user-friendly interface with user-controls and further technical improvements to be included.
7- References

4-Radboud Winkels, Jelle De Ruyter and Henrky Kroese; Determining Authority of Dutch Case Law; Legal Knowledge and Information Systems, Vol. 235, pp. 103-112; 2011.
5-T. J. M. Bench-Capon, P. E. S. Dunne & G. Stanford (1997); Modelling Legal Documents as Graphs, Information & Communications Technology Law, 6:2, 103-120.
6-Mohammad Hamdaqa, Abdelwahab Hamou-Lhadj; An approach based on citation analysis to support effective handling of regulatory compliance; Future Generation Computer Systems 27 (2011) 395-410.
7-Airi Salminen, Frank Tompa; Communicating with XML; Springer Science+Business Media, 2011.
13-Cliff Binstock, Dave Peterson, Mitchell Smith, Mike Wooding, Chris Dix, Chris Galtenberg; The XML Schema Complete Reference; Addison Wesley, 2002.
16-Elliotte R. Harold; XOM Design Principles; Proceedings of the Extreme Markup Languages-2004 Conference; Montréal, Québec, Canada.
20-Emile de Maat; Making Sense of Legal Texts; SIKS Dissertation Series No. 12-26; Ridderprint BV, 2012
21-Michel Chein Marie-Laure Mugnier; Graph-based Knowledge Representation, Computational Foundations of Conceptual Graphs; Springer-Verlag, 2009.
Finding and Visualizing Context in Dutch Legislation

Radboud WINKELS and Alexander BOER
Leibniz Center for Law, University of Amsterdam, The Netherlands

Abstract
This paper describes preliminary research on automatically determining relevant context to display to a user of a legislative portal given the article they are retrieving, purely based on ‘objective’ criteria inferred from the network of sources of law. A first prototype doing this and visualizing the result in a coloured graph with variable nodes and edges is presented and a first formative evaluation by one legal expert user. Results are promising, but there is plenty of room for improvement.

Keywords. network analysis, legislative portal, MetaLex, references, impact analysis

Introduction

The official Dutch portal for national legislation (www.wetten.nl) 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 1). 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 1. Interface on wetten.nl: On the left you can browse through all the chapters of the regulation and in the text of article 35 on the right there are 2 outgoing references visible as hyperlinks.

This paper describes ongoing research to improve the user interface on Dutch legislation by providing more and broader context to users for the article(s) they are inves-

1Corresponding Author: E-mail: winkels@uva.nl.
tivating. We try to exploit the network structure that laws exhibit. Articles refer to other articles, in the same law or in other sources. This forms part of the network. Besides, earlier versions (or later, if one examines laws from the past) of the document in focus are related and some others might be through delegation of power.

Our research question is: given a particular document (article) in focus, can we determine other relevant documents purely on the basis of ‘objective’ meta-information? That is, we do not look at or interpret the content of the documents, nor do we use metadata added by other sources than the official owners and publishers of the documents.

Network analysis in a legal context has been done before, both on legislation (e.g. [5][1]) and on case law (e.g. [2][8]). The work on legislation however focused on analysing the entire network at the level of laws [5] or within one specific law [1]. We are looking for a small local network, given one starting point, possibly encompassing several laws or other sources.

We will first explain how we construct a context network from the legislation given a particular document in focus. Next we will present a first formative evaluation of the context provided by an expert user and we end with some conclusions and suggestions for further work.

1. Creating a Context Network of Law

The “Metalex Document Server” (MDS) contains all regulations from the Dutch portal in CEN MetaLex XML [2][3] and as RDF linked data ([3]). This format is much more suitable for our purposes than that provided by the official portal. MDS currently contains more than 280 million RDF triples (April 2013), and this number is growing every day since every change to the wetten.nl site is added to the triplestore.

Regulations can be identified by their BWB-ID, e.g. the foreigners law has ID “BWBR0011823”. This ID can also be used to find a document at the wetten.nl site by entering it in the url; the entire foreigners law can be found at: wetten.overheid.nl/BWBR0011823. ‘BWBR’ stands for “Basiswettenbestand” (basic law file), the content management system for all Dutch regulations that underlies the portal. An ‘R’ following ‘BWB’ indicates that the document is a regulation, a ‘V’ indicates a treaty (‘verdrag’). The 7-digit number does not carry a specific meaning. The opaqueness of the BWB identifier is unfortunate, but hard to avoid, as the title of a regulation may change over time and cannot be used.

CEN MetaLex distinguishes the source of law as a published work from its set of expressions over time, and the expression from its various manifestations, and the various locatable items that exemplify these manifestations, as recommended by the IFLA Functional Requirements for Bibliographic Records (cf. [6]). References in legislation go from a specific version (expression) to the work level of the entity it refers to. Suppose article 35 of the foreigners law of September 1st 2010 refers to article 37 of that same law, then the reference leads from expression 2010-09-01 to the work art. 37 (see figure 2).
Figure 2. Expressions of works can reference other works. Sometimes new versions of citing and cited works
are created at the same time, in which case the expression 'virtually' cites an expression, because the work it
cites has only one valid expression during its life time. This is the case for the expressions of citing article 37
and cited article 108 in this example.

Incoming References

To create a network from an article in focus we start by collecting all incoming references. We can use the SPARQL endpoint of MDS for this. Since all references to our focus node refer to the work level and we know the work level of the article in focus, this is straight forward.

To find the distinct incoming references ('cites' in MetaLex) of the focus node article 35 of the foreigners law we can use the following SPARQL query:

```
SELECT DISTINCT ?s WHERE {
  { ?s <http://www.metalex.eu/schema/1.0#cites> <http://doc.metalex.eu/id/BWBR0011823/artikel/35> } }
```

The result will be something like this:

```
<?xml version="1.0"?>
<sparql xmlns="http://www.w3.org/2005/sparql-results#">
  <head>
    <variable name="s"/>
  </head>
  <results>
    <result>
      <binding name="s" uri="http://doc.metalex.eu/id/BWBR0012288/circulaire.divisie/8/circulaire.divisie/3/tekst/2/a1/1/extref/1/2012-05-05"/>
    </result>
    <...more results...>
  </results>
</sparql>
```

4 http://doc.metalex.eu:8000/test/
The incoming references (bound to the variable ‘s’) contain the part-of structure of the referring document. E.g. from the example above one can obtain the information that the reference to the article in focus comes from the first paragraph of subsection 3 of section 8 of the foreigners circular letter (BWBR0012288: “Vreemdelingencirculaire 2000 (C)”) that was valid on May 5th 2012. The “extref” indicates that it is an external reference as opposed to internal ones.

**Outgoing References**

Finding the outgoing references is a little trickier, because it are always expressions that make these references, as explained above. In other words, we have to check all expressions of the article in focus for outgoing citations. The following SPARQL query finds all the outgoing references and filters them for the focus node.

```
SELECT DISTINCT ?s ?o WHERE {
  { ?s <http://www.metalex.eu/schema/1.0#cites> ?o }
  FILTER regex(str(?s),"BWBR0011823.*artikel.35")
}
```

The result will be something like this:

```
<?xml version="1.0"?>
<sparql xmlns="http://www.w3.org/2005/sparql-results#">
  <head>
    <variable name="s"/>
    <variable name="o"/>
  </head>
  <results>
    <result>
      <binding name="s"> http://doc.metalex.eu/id/BWBR0011823/hoofdstuk/3/afdeling/4/paragraaf/2/artikel/35/lid/1/lijst/3/1i/2/a1/2/extref/1/n1/2013-01-01 </binding>
      <binding name="o"> http://doc.metalex.eu/id/BWBR0001854/artikel/37a </binding>
    </result>
  </results>
</sparql>
```

The variable ‘s’ will be bound to all expressions of the article in focus (in this case still art. 35 of the foreigners law) that contain a reference to some other source of law that will be bound to the variable ‘o’. These are cited at the ‘work’ level - in the example above article 37a of the Dutch penal code (“Wetboek van strafrecht”).

---

5 A softlimit of -1 is entered in the SPARQL endpoint to get all the results.
Besides the incoming and outgoing references of the focus node, we decided to also include the incoming references of the nodes the article in focus refers to (given by the outgoing references) in the context network. This can be achieved by the same type of SPARQL query as used for finding the incoming references of the focus node.

**Seeding the Context Network**

The idea is to provide a user of the legislative portal with a context given a particular node in focus. As we do not have an actual working system yet, and therefore no users that we try to help, we had to find starting points for our context creation. A legal expert at the Dutch Immigration and Naturalisation Service (IND) indicated that the articles 35 and 16 of the foreigners law ("Vreemdelingenwet") were good starting points.

Not all articles in the network made from a starting node are potentially interesting or can even be displayed. Some articles have hundreds of incoming references from different expressions, so a selection has to be made.

2. **Analysing the Context Network of Law**

There are several ways one could determine the potential importance of ‘nodes’ in the context network of a focus node. We will discuss some.

**Superiority.** There is a hierarchy in the legal system that is connected to delegations of power. Laws can delegate further specifications to lower types of regulations, e.g. policies or decrees. The main law does not - and most often cannot - refer to these lower regulations, but the opposite does happen. The lower regulations will refer to the main law to identify its source of power and may further refer to it to fill in specific concepts or norms from the main law.

Whether an article is from a lower regulation or not, cannot be derived from the BWB number, but because we know that laws do not refer ‘downwards’, we can hypothesize that outgoing references are more important than incoming ones.

**Ex- or internal.** We can distinguish external from internal references. External references point to (parts of) another source of law. An internal reference is a reference that points to an article in the same regulation. It is difficult to say whether external references are more important than internal ones, or the other way around. On the one hand a user may not think about other sources and a reminder may be important. On the other hand the user may be unfamiliar with the current law and internal references may be more important. It is probably safe to assume that references from an article to parts of itself are not important. We will have to experiment with assigning weights to external and internal references.

**Definitions.** The internal references can point backwards or forwards (anaphora and cataphora respectively). In general the anaphora are there to use a definition that is given in a previous article. For this reason they are considered interesting. The cataphora are usually there to introduce what will be described in the following sections and are therefore considered less important.

---

6Because the lower regulation does not exist yet at the moment of drafting the law.
Dynamics. The recency of a change to an article can indicate that it is important, assuming that users may (still) be unaware of the change. The total number of expressions that exist of one work may also indicate importance, i.e. such a work changes a lot.

Centrality. A traditional network measure. The number of incoming references to a certain node is a way of sorting nodes on importance (in-degree). We expect that the more incoming references a node has, the more important it is.

Based on these features we can compute the weight of nodes in the context network. We will compute these in several ways and have users tell us which are more useful than others by visualising the resulting networks with the weight expressed in the size of the nodes and edges.

Before anything can be calculated the XML files containing the incoming and outgoing references are read and sorted into arrays. After this some filtering with regular expressions takes place to find out for example whether the references are internal or external. Then we compute overall weights and visualise the network.

3. Visualisation

For the visualisation the ‘JavaScript InfoVis Toolkit’ is used[7]. In its ForceDirected graph the nodes and edges can all be modified and information can be displayed upon selecting a node. Animations are also possible, which will be useful for showing the changes over time to the network. The visualisation toolkit takes JavaScript Object Notation (JSON) as input where the id, name, adjacencies and visualisation information of the node should be specified. With the size and colour of the nodes and edges the importance of that particular node can easily be conveyed.

Figure [8] shows an example of our prototype visualising the context network for article 35 of the foreigners law. One can start the system with any seed node. The sizes of the nodes show how ‘relevant’ those nodes are to the focus node; the bigger the more relevant. Different colours are used to indicate the different laws the nodes are in. The edge width currently shows how many previous versions of the expression also refer to the connected node. Outgoing references are visualised with square nodes and incoming references have triangle shaped nodes[8]. The information displayed is limited to the 30 most important incoming nodes and the 10 most important outgoing nodes.

All the dates at which something changes in the visualisation are shown on the left hand side as buttons and when pressed the visualisation changes to display the network information from the beginning up to that date.

Finally, the prototype makes it possible to display the current weight settings and to alter these.


[8] It would be more natural to use arrows, but due to a bug in the toolkit these either pointed in the wrong direction or were not displayed at all.
4. A Simple Formative Evaluation

Ideally we would have at least 30 experts that work daily with the legislative portal in different domains of law, to determine the importance of articles that are connected to the node in focus. Given the unfinished status of our prototype we asked only one expert of the IND in a formative evaluation to give feedback on the displayed connections in the two cases of our seed nodes. This feedback may lead to new insights in variables that can also be derived from the network.

The people of the IND already use mind maps to visualise relevant information. These mind maps are shared among groups at work, but it costs a lot of time to create them. Concepts and articles are linked in those mind maps and articles of the same work have the same colour. The mind maps are one of the tools they use to assess the impact of changes in the law on other parts of the law. All relevant information is now retrieved from the portals www.wetten.nl and www.migratierecht.nl This last portal only contains those parts of the law and other documents that are relevant to immigration law.

Our ForceDirected visualisation is similar to these mind maps since it also connects articles with each other. The differences are that there is no importance attached to the articles in the mind maps and that our visualisation does not contain general concepts.

For this first test, different visualisations of our context networks for the two seed articles 35 and 16 of the foreigners law were built and printed out on cards and the expert was asked to sort them in order of usefulness. We hoped the ranking and the feedback would help us improve our network building heuristics.

4.1. Results

These were the main comments of our expert:
The people working at the IND concerned with impact assessment of legislative changes need an overview of all incoming and outgoing references of affected articles. It is not possible to easily find the incoming references on either of the portals they use, so this feature of our visualisation and general functionality of the MetaLex document server will definitely be useful to them.

Incoming and outgoing references are equally important in their line of work, because when an article changes all the connected nodes will be influenced and it does not matter whether they are incoming or outgoing.

It differs per focus node whether anaphoric or cataphoric references are more important and it is best if a user can set this weight if it is explained what it does exactly. In the case of article 16 of the foreigners law, it refers backwards to article 14 which is the article that gives the minister the ability to give or deny people the right to live in the Netherlands. The reference forwards to article 17 names the special cases. The last one is probably more important in this case than the anaphoric reference.

The only thing that really influences how important a reference is, is its position in the hierarchy; so laws are more important than decrees or policies.

With the current context network about 95% of the relevant articles are found and the last 5% are internal documents or work instructions that can not be found on the official legislative portal.

Going levels deeper and finding the incoming or outgoing references of the found connections to the focus node, does not result in a better network. It is more information than one actually needs. It might be that this is only irrelevant for this domain.

For the visualisation of the found network it is best to have both the visualisation in the ForceDirected graph and text based results. The colouring of the nodes per work is a good way to visualise them since this is also done in the mind maps. What is missing are the titles of the works because now it only says for example “BWBR0011823/artikel/33” and a link to the actual text would also be very useful.

The articles connected to the two seed articles were also ranked from important to less important by the expert. Table 1 gives the ranking for article 35 of the foreigners law. The difference in ranking is based solely on the hierarchy. Many of the weights that can be set are therefore not used and the only important change was setting the weight of outgoing references the same as for incoming.

5. Conclusions and Future Work

The question we set out with was whether it would be possible to determine which other legal documents to show a user of a legislative portal that has a particular document (article) in focus, purely on the basis of ‘objective’ network information. Our first answer is “yes”, but there is plenty of room for improvement.

The context networks we generated for two starting nodes in the domain of immigration law contained all of the relevant documents, according to our expert from the IND. The few we missed were not available in the portal we used to produce our net-
The expert evaluating our context networks was focusing on the task of impact analysis after legislative changes. This may have led to other context needs than for instance deciding in a particular immigration case. The fact that she considered outgoing links as important as incoming could be an example. This too needs to be further investigated.

She also missed the titles of the nodes in the visualisation. The visualisation works for an immediate overview of what nodes are important, but it would be better to combine it with a text based visualisation. In a text based visualisation one can also show the difference in importance using the size of the title and how much of the text is displayed. It can be compared to the look of a newspaper that also uses different headings to convey the importance of a piece of text. We need to look into research on automatic pagination (see e.g. [4]).

For the ‘time traveling’ aspect of the visualisation we would prefer something like a ‘time slider’ in the future.

For deciding on what to put in the context network and the relative importance of nodes and edges, we also want to investigate some other features:

- Look into the content of documents to e.g. extract concepts [7], so one could see whether an article contains special cases or whether it gives the minister the power to do certain actions. These concepts can help in determining the importance of documents.
- In some sections of the law there are references that say that a certain thing does not hold for a specific article. This is a reference, but perhaps it needs to be treated differently from a ‘positive’ reference.
- Including networks of other sources of law: case law and legal doctrine. Case law and legal doctrine also refer to legislation and this could be used to determine the importance of articles. Moreover, related cases and doctrine could be presented in the context network. The in-degree coming from cases or doctrine to an article needs to be handled differently than those coming from legislation. Case law always refers to a specific version of the legislation, which does not have to be the latest. With doctrine it depends. It may refer to specific expressions, to the work

<table>
<thead>
<tr>
<th>Connection</th>
<th>Ranking by expert</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreigners Law 2000 article 33</td>
<td>1</td>
</tr>
<tr>
<td>Penal Code article 37a</td>
<td>1</td>
</tr>
<tr>
<td>Foreigners Decree 2000 header</td>
<td>2</td>
</tr>
<tr>
<td>Foreigners Circular 2000 (A) Chapter 3 article 7.3.2</td>
<td>3</td>
</tr>
<tr>
<td>Foreigners Circular 2000 (C) Chapter 8 article 4</td>
<td>3</td>
</tr>
<tr>
<td>Foreigners Circular 2000 (C) Chapter 8 article 3</td>
<td>3</td>
</tr>
<tr>
<td>Foreigners Circular 2000 (C) Chapter 8 article 2</td>
<td>3</td>
</tr>
<tr>
<td>Foreigners Circular 2000 (C) Chapter 8 article 1</td>
<td>3</td>
</tr>
<tr>
<td>Foreigners Circular 2000 (C) Chapter 21 article 4</td>
<td>3</td>
</tr>
<tr>
<td>Foreigners Circular 2000 (C) Chapter 8 article 5</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 1. The connections to article 35 of the foreigners law with importance rankings given by the expert from 1-3 with 1 being highest.
level, or even to future expressions or ones that have never existed or will never exist.

Acknowledgments

We would like to thank Anna Keune for designing and implementing the first version of the prototype and Christiane Buschman and Maureen Bredewold of the IND for helping with the network seeds and evaluating the prototype.

References