Making sense of legal texts

de Maat, E.

3 References

References are the counterpart of document structure. All sources of law are glued together by means of references, and their position in the legal framework is also made clear through references. The next step in our proposed modelling process, depicted in figure 4, is to detect and annotate any references found in a legislative document. This step starts out with a legislative document in which the structure of the text has been made explicit, and annotates any references in that document. This means marking any text that forms a reference, and adding the target to that reference.

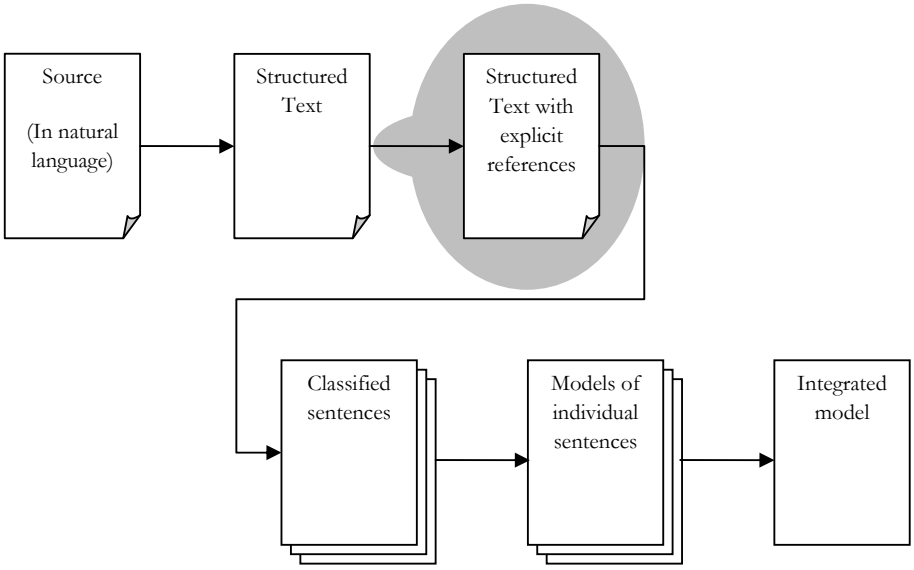

Figure 4: Parsing references step

Legal texts refer to other legal texts (which may be part of the same document) for a variety of reasons. First of all, lower regulations that are promulgated under the authority of a higher regulation will refer to that regulation to establish their legal ground[45]. Next, regulations that change other regulations need to refer to these regulations in order to make clear where the changes are applied.

Within the rules themselves, references are used for a variety of reasons, such as:

- To set the scope for a definition (indicating that the definition is valid throughout that scope): *for the application of article 12, it is understood by …;*
- To re-use definitions and descriptions that have been made in other texts: *one of the persons meant in sub 2, item a;*
- To indicate that certain rules only apply if other rules have already been applied: *if article 14 is applied*;
- To indicate that a certain rule is an exception to another rule, and that the other rule does not apply in this case: *in exception of article 17* or *article 17 does not apply;*

[45] Such a higher regulation will have delegated power to a lower regulation by means of a phrase like *By royal decree, specific rules will be set to* .... In a sense, this is a reference to a future regulation (the royal decree). An important difference with the other references is that these references cannot be resolved based on the information found in the source document, but require the target document (with it reference back) in order to be resolved. These references will not be discussed further in this thesis.

- To explicitly indicate that a rule is not an exception to another rule: *Without prejudice to article 17* or *article 17 does apply*.

Other documents may include other purposes of references; case law, for example, will indicate which rules have applied and which have been deemed not applicable. Linked together by means of these references, the sources of law form a network.

In Boer et al. (2009), different bibliographic entities are discussed that may be the target of a reference. These are inspired by the functional requirements for bibliographic records (International Federation of Library Associations and Institutions, 1998). The bibliographic entities are:

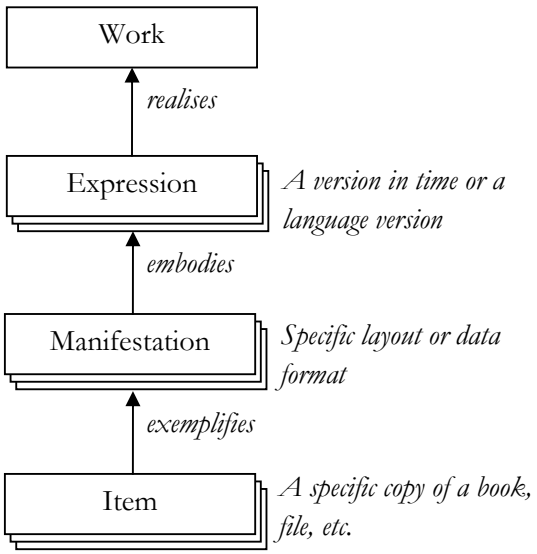

Figure 5: Bibliographic entities

- A *work*, which is some regulation, e.g. the Coin Act 2002.
- An *expression*, which is a specific version of a regulation. This could be the original version (e.g. the Coin Act 2002 as it was originally published) or some later version (e.g. the Coin Act 2002 as modified on November 22nd, 2006. Each work has at least one expression; many works will have only one. An expression is said to *realise* a work.
- A *manifestation*, which is an expression in a specific formatting and layout, e.g. the PDF created by SDU of the Coin Act 2002 as modified on March 1st, 2007. A manifestation is said to embody an expression.
- An *item*, which is a specific instance of a manifestation, i.e. the copy of the aforementioned PDF file that resides in the documents folder of my computer. An item is said to exemplify a manifestation.

Legislation refers to a work. When applying the legislation, all references should be seen as references to the current version of that work (or, when applying the legislation on something that happened in the past, all references should be seen as references to the version that applied at the time being considered). Even the references in amending laws refer to a work; the difference with a regular law is that it will only be applied once. However, at the time the amending law is made, it is unknown to what expression it will be applied, and the textual references are made at the level of a work. Case law, on the other hand, usually refers to a specific version of a work (i.e. the expression that was relevant for the case). It may even happen that they refer to page numbers, in which case they refer to a manifestation. Strictly speaking, doctrine also refers to a specific version, though it may be valid through a specific range of versions, and could be seen as referring to a specific subset of expressions (whereas a work encompasses all expressions). When a new version of a law appears, doctrine may also be upgraded, leading to a new version of the doctrine to go with the new version of the law.

Many existing applications and formats for legal content have an option to mark references inside the text; often, these take the form of HTML hyperlinks or something similar. These are very generic, and do not give us much insight into the structure of the references.

Also, in existing systems, this can lead to problems when the hyperlink leads to a specific item (instead of an expression). This is often the case when a collection comes from one publisher, who has added hyperlinks to other files inside the collection. Such a collection will not connect to items outside the collection and material from other publishers. In this case, it may be useful to detect the references anew to create a complete network that is no longer confined to the collection (see Winkels et al., 2005).

Like the headings that indicate the structures, references follow fixed format, obviously facilitated by the fact that the documents being referred to all have a similar structure. These structures enable the automated detection of references. After a reference has been detected, it should be resolved, meaning establishing the identity of the work or expression that is being referred to.

3.1 References to Documents

Laws and treaties are often referred to by their short title, such as the *Customs Act*[46] or *Treaty establishing the European Economic Community*[47]. If no short title exists, the guidelines prescribe a format including the date, subject and publication number, such as:

the Act of November 4th, 1950 for the establishment of additional regulations regarding military pensions that have been given out during hostile occupation, as well as modification of several laws, which provide rules regarding military personnel (Stb. 1950, K 479)[48]

Often, this format is somewhat modified, leaving out the subject:

the Act of November 4th, 1950 (Stb. 1950, K 479)

For treaties, the guidelines prescribe a similar format, including date, place, subject and publication number[49].

the on June 20th, 1956, in New York negotiated Convention on the Recovery Abroad of Maintenance (Trb. 1957, 121)

Finally, there is also a prescribed format for European directives:

---

[46] Douanewet

[47] Verdrag tot oprichting van de Europese Gemeenschap

[48] de Wet van 4 november 1950 tot nadere vaststelling van de regelingen op het gebeid van militaire pensioenen, welke gedurende de vijandelijke bezetting zijn uitgevaardigd, zomede nadere wijziging van verschillende wetten, welke regelen geven inzake militair personeel (Stb. 1950, K 479) (Verwijzing uit de Wet privatisering ABP, artikel 77)

[49] **Aanwijzing 88, eerste lid**
Een verdrag wordt aangehaald overeenkomstig het volgende voorbeeld:
het op 20 juni 1956 te New York tot stand gekomen Verdrag inzake verhaal in het buitenland van uitkeringen tot onderhoud (Trb. 1957, 121).

Directive 1999/2/EC of the European Parliament and of the Council of 22 February 1999 on the approximation of the laws of the Member States concerning foods and food ingredients treated with ionising radiation (PbEG L 66).

In addition to references by full title or short title, some more formats are used. A law can refer to itself using the anaphor *this law*. Sometimes, other anaphors like *that* law or *the law mentioned before* are also used. Furthermore, a law can define an abbreviation for a document (often *the law* or *the treaty*) which can then also be used as a reference. For example:

Immovable Property Valuation Act, article 2, definition of the law[50]
In this law, it is understood by the law: the Immovable Property Valuation Act.

So, in this law, any reference to *the law* refers to the Immovable Property Valuation Act (which happens to be the law itself).

3.2 References to Parts of Documents

A part of a document, such as a chapter or article, is referred to by means of its category or index (as discussed in section 2.1 and 2.2), e.g. *chapter II* or *article 24*. The category *article* is sometimes abbreviated to *art.*: *art. 24*. Subparagraphs and list items do not include a category in the header, but they do use a category in a reference, e.g. *item a*[51]. For numbered subparagraphs, it is common to use an ordinal instead of the actual index, e.g. *first subparagraph* instead of *subparagraph 1*.

In order to refer to a part of a specific document, the reference to the part is combined with the reference to the document: *article 2 of the Mining Act*. Likewise, a reference to a subpart can be combined with the reference to a containing part: *article 2, first subparagraph* (which again can be combined with the reference to the containing document).

For such a "layered reference", there are three ways to arrange the layers:

- Zooming in: The reference starts at the top level and travels down to the lowest level: *Mining act, article 2, first subparagraph*.
- Zooming out: The reference starts at the lowest level and travels up to the top level: *first subparagraph, article 2, Mining Act*. Between two levels, the word *of* can appear, so *first subparagraph of article 2 of the Mining Act* or *first subparagraph, article 2 of the Mining Act* can also occur.
- Zooming in, then zooming out: The reference starts at some level (usually the article) then "zooms in" and at the end "zooms out" again: *article 2, first subparagraph, of the Mining Act*. The "zooming out" part usually consists of one step, sometimes two, but seldom more.

It is also possible for a reference to refer to more than one part of the law. This is most often done by providing two or more indices. The category can be singular or plural (i.e. both *article* and *articles* are used in a reference to multiple articles). Some examples: *articles 12 and 13*, *article*

[50] **Wet waardering onroerende zaken, artikel 2, aanhef en negende onderdeel**
In deze wet wordt verstaan onder de wet: de Wet waardering onroerende zaken.
[51] In fact, there are several different categories used to denote list items. The guidelines prescribe the use of *onderdeel*, but older laws also use *onder*, *sub* or *letter*.

14, 15 and 18, *item a, b and f*. For articles, the abbreviation *art.* is again used, as well as the plural *artt.*

Alternatively, a range of indices is specified by specifying the first and the last index of the range. They are either separated by means of a hyphen (though this is discouraged in the guidelines[52]) or the words *up to and including*: *article 12 - 15, items a - f, articles 12 up to and including 15*.

As with references to a single part of a document, references to multiple parts can be combined with references to containing parts of documents: *first and second subparagraph, article 2 of the Mining Ac* .References to different levels can also be combined *articles 7, second subparagraph and 8 up to and including 12*.

If the reference contains a lot of repetition, because the same numbered subparagraphs or items from several articles are referenced, then this can be shortened by *each time*. For example, a reference to *articles 8d, first subparagraph, 53d, first subparagraph and 133d, first subparagraph* can be written as *articles 8d, 53d and 133d, each time the first subparagraph*.

It is also possible that a reference contains exceptions on a given range or super part: *chapter 3, except article 17*.

Lists have an introduction and, optionally, a conclusion, which are referred to as *introduction* and *conclusion*. They are seldom referred to on their own; usually they are quoted together with some of the list items: *article 12, introduction and items i and j*.

Finally, like full documents, parts of documents are sometimes referred to using anaphors like *this article* or *the previous article*.

In appendix B, it is shown how these references can be detected using patterns in GATE.

3.3 Resolving References

After we have found a reference, it should be resolved, meaning that the identity of the document being referred to is determined. This identity will usually take the form of a Universal Resource Identifier (URI). Such an identifier can be meaningful or opaque.

If the identifier is meaningful, then the identifier is somehow linked to its meaning. For example, within the NIR project, it is prescribed that the identifier for *article 2* should be *art2*, and that the complete identifier for *article 50 of the Customs Act* should be *#art2* appended to the URI for the Customs Act (see Spinosa, 2001). Using patterns, references and parts of references can be identified, so it is possible to recognise the text *article 50, first subparagraph of the Customs Act* as a reference that refers to *subparagraph 1* of *article 50* of the *Customs Act*. Using the NIR rules, we can use this information to create the correct URI.

In a system that uses identifiers that relate directly to the names or indexes of the elements of the document, it is easy to determine the right URI. However, there are some disadvantages to such a system. First of all, it is does occur that laws have two articles that have the same

---

[52] **Aanwijzing 65**
Het einde van een periode of reeks wordt aangeduid met de uitdrukking "tot en met".

number, or two subparagraphs within an article to have the same number, etc.[53] In such cases, a meaningful identifier is still possible, but will have to rely on more than just the names and indexes by including position: the *first article 12* and the *second article 12*. A second problem occurs when a part gets renumbered or renamed. In such cases, the identifier should change too, which means that some metadata is needed to link the part to its previous incarnations.

So, many systems use meaningless identifiers, which have the obvious disadvantage that the identifier does not relate in some way to the name or index of the text. This means that no URI can be constructed; it must be retrieved from some list.

It is also possible that the identifier is partly meaningful, partly meaningless, in which case a part of the URI can be constructed, and the other part needs to be retrieved.

The method described above only works when a reference is complete; that is, it contains the identity of the document that is being referred to. Many references encountered in a law do not contain that information. In such cases, we need to determine the work being referred to before we can determine the correct URI.

In most cases, an incomplete reference to a document part refers to that part within the same document (or within the same part of a document). So:

- A reference to *article 72* is a reference to article 72 of the law it is encountered in.
- A reference to *subparagraph 2* is a reference to subparagraph 2 of the article of the law it is encountered in.

A common exception to this is found in sentences that describe changes in existing laws. Such sentences are often preceded by a scope declaration (see section 4.4.3). Such a scope declaration sets the location of any changes being made. For example:

Customs Act Introduction Act, article XIX[54]
To the Aviation Act, the following changes are made.

Any incomplete references following this scope declaration do not refer to the containing document but instead to the document being set in the scope declaration. So, if this scope declaration is being followed by a subparagraph that refers to *article 37a*, then this is a reference to article 37a of the Aviation Act. This does not apply to any reference contained in quoted text or text to be inserted, as references in those texts should be resolved as part of the document that they are quoted from or are to be inserted in. In order to correctly resolve these references, these scope definitions need to be known.

Another group of references that do not contain the identity of the document that is being referred to are the anaphors. There are a number of different groups of anaphors that may be encountered.

---

[53] Such mistakes are seldom present in a new law, but are usually introduced when changes are made later on.
[54] **Invoeringswet Douanewet, artikel XIX**
In de Luchtvaartwet worden de volgende wijzigingen aangebracht.

1. References to *this law*, *this article*, etc.: These references refer to the law (or the article, etc.) in which they are found, and are easily resolved if the structure and identity of the document are known.
2. References to *that law*, *that article*, etc.: These references refer to an earlier reference (most likely the previous). Providing that previous reference has been found, all the information needed to resolve such a reference is available.
3. Reference to *the previous article*, *the previous subparagraph*, etc.: References to a previous structure part are easily resolved if the structure and identity of the document are known.
4. Nested references, such as *the articles mentioned in the previous article* are a reference to a reference that can be found in a specific location. This requires that the references to that location (in the example: *the previous article*) is first resolved. Provided that the reference in that location has already been found, the complete reference can now be resolved.

3.4 Experiment

In order to test this approach, a parser for recognising references to laws has been constructed based on the patterns described above, excluding the patterns for exceptions and repetitions. These results have been published before in de Maat, Winkels and van Engers (2006). Similar systems have been built before for other jurisdictions. Bolioli, Dini, Mercatali, and Romano (2002) and Palmirani, Brighi and Massini (2003) have created reference detection systems for Italian law. The system built by Palmirani et al. was capable of (partially) detecting 93.6% of all references in a set of Italian IT laws. Martínez-González, de la Fuente and Vicente (2005) have created a system for Spanish legal texts, which had a recall of 54%.

The Dutch parser was applied to six randomly selected Dutch laws[55]. These were selected to include one law from before 1900 and another law from between 1900 and 1950, since the language used in those older texts may differ from the language used in modern text.

The results are presented in table 4. The references have been split into two groups: Simple references, which are non-layered references to a single element, and complex references, which are references that are layered and/or refer to multiple elements. A complex reference was considered to be partly found if part of the text was recognised as a reference, but the complete reference was not recognised. The skipped column shows references to documents other than laws, which were not included in this experiment.

[55] In addition to this formal test, the parser has also been informally tested on several other laws and decrees, in which it performed very well, with few errors.

		Simple		Complex				
		Found	**Missed**	**Found**	**Partly**	**Missed**	**Skipped**	**False**
Wet tarieven in burgelijke zaken	1843	24	0	27	5	0	1	0
Natuurschoonwet	1928	40	1	46	1	0	0	0
Wet aansprakelijkheid olietankers	1975	38	0	35	1	0	10	1
Wet op de lijkbezorging	1991	69	0	47	2	0	0	0
Wet gemeentelijke basisadministratie persoonsgegevens	1994	251	1	156	8	0	6	2
Wet op het notarisambt	1999	118	1	127	7	0	0	4
		540 (99%)	3 (1%)	438 (95%)	24 (5%)	0 (0%)	11	7

Table 4: Test results for detecting references

The parser achieved good results, finding 99% of all simple references, and 95% of all complex references. The few misses were caused by missing labels, names or patterns from the parser. If such labels and patterns occur more often, they can be included as well[56] (though there will always remain some patterns that are too rare to include).

False positives occur if one of the labels used for the categories is used in a different meaning. For example, a subparagraph is called *lid* in Dutch (which also means *member*). Hence, confusion can occur when the text uses the text *the first member*, it may be a reference to either the first subparagraph or the first member of e.g. a committee. Such a false positive can sometimes be detected during the resolving of the reference, if the perceived target does not exist.

An adapted version of the parser has been used by Opsomer et al. (2009), who expanded the parser for Flemish legal texts. It was used on 1600 text fragments, ranging in size from very small (part of an article) to large (an entire chapter). Each fragment was part of Flemish, Belgian or European environmental and energy legislation, all written in Dutch. After the parser had recognised a reference, it was resolved using a module that searches a legislation database (EMIS Navigator, a database with environmental legislation applicable in Flanders) for the text being referenced. This procedure was also used to identify false positives: if a (perceived) reference referred to a non-existing text, it was deemed to be false. Opsomer et al. estimate that the parser found 85% of all references, with a precision of 95%.

3.5 Generalisation to Other Sources

The method described above works well for finding references in law texts to other laws, treaties and European directives.

Other (national) regulations follow the same writing style as laws, and are subject to the same guidelines. As such, the same patterns used to find references in laws will be effective in finding references in regulations. However, laws refer only to other laws, treaties and European legislation, whereas lower level legislation will also refer to other lower level legislation. For those references, the patterns will have to be expanded. Just as for the

[56] The missing names have been included in any case; the parser used a list based on current laws, and some of the references found referred to retracted laws.

documents already described above, the Guidelines for Legal Drafting prescribe a format for references to lower level legislation, which can be used a basis for expansion of the patterns.

In addition to lower level legislation, commentaries and case law form an important source of legal information as well. These documents do not have to abide by the official guidelines, and hence are less uniform. Van Opijnen (2010) found that when referring to secondary EU legislation, Dutch judgments employed a great variety of formats (he gives examples of ten different formats used). Still, he found that a number of patterns would suffice to detect the references in all these formats. In general, the writing style used for these references remains close to the style used in legislation, and as such, the patterns used to find references in legislation will likely form a good starting point for finding references in jurisprudence and commentaries. Some expansions will be needed:

1. Patterns for references to new categories: Regulations do not refer to case law and commentaries, but case law and commentaries do contain such references. Patterns will be needed to cover these additional documents.
2. Unofficial names: Commentaries and jurisprudence will often refer to documents using an unofficial name. Just as official names, such names are difficult to detect using patterns, and should be collected in a list, to be used in the same way as the list of official names.
3. Variant patterns: As Van Opijnen found, commentaries and jurisprudence are somewhat less strict, and apply more different formats; hence, more patterns are likely necessary in addition to those already used for laws.

Next to these difficulties in detecting references in the text of commentaries and case law, there is also an additional issue with regards to the resolving of those references.

As mentioned above, regulations refer to a work. As such, we only need to know the identity of the document being referred to in order to resolve the reference. Commentaries and jurisprudence refer to a specific expression of a work[57]. Ideally, to resolve the reference, we need to know the identity of the document but also the version of the document. Commentaries will often refer to the version of a regulation as it was valid when the commentary was published, or, alternatively, they refer to some future version of the regulation. Case law usually refers to that version of the law that was applicable to the facts being judged. In both cases, the text will contain some clues (dates) relating to the relevant version[58]. The question is whether it is possible to automatically extract these dates from the documents.

3.6 Conclusion

References in Dutch laws are very well structured, and can be easily detected using patterns. This is confirmed by a test on six very diverse Dutch laws, which showed an accuracy of 97% and barely any false positives. A similar test on a more diverse Flemish corpus, which included

[57] Though it is quite possible that a commentary is applicable to a certain range of expressions, or even to the entire work.

[58] In addition to the date of the version, the date of publication of the referring document may also be relevant. The future version of a law of January 1st, 2012 that is being referred to in a commentary may very well be a different version than the actual version of that law on January 1st, 2012.

documents from different jurisdictions, gave an accuracy of 85%, with 95% precision. Furthermore, such parsing gives us most of the information needed to resolve the references, though for a complete approach it is also necessary to scan the law for definitions defining abbreviations of document names and for scope definitions.

The study of the structure of references and the different forms that references can take has led to the identification of some less-common structures. These structures have been included in the European XML standard for marking up legal sources, which includes mark-up for exception constructions and each-time constructions. Such support was not present in the predecessors of CEN/MetaLex, meaning that these references could not properly be annotated.

As the results mentioned show, there are a few troubles with the recognition of references. The variety of indexes that are used do increase the number of rules needed to parse everything, but it does significantly increase the difficult of the task. But when it comes to resolving references, those references that do not explicitly refer to their target(s) by means of an index pose something of a problem.

The first type of such references is the range. When a range is referenced, we can identify the start and the end of the range, but this does not give us full knowledge of what parts of the document are referenced. For example, a reference to articles 12 to 15 could be, depending on the structure of the target document:

- a reference to articles 12, 13, 14 and 15;
- a reference to articles 12, 13, 13a, 14 and 15;
- a reference to articles 12, 13 and 15.

This set of targets may differ from one version of a law to another version of the law. This means that for each version, we need to calculate the exact set of items referred, if that information is required for our purpose.

A related question that arises is: Is this effect intended by the legislator? When he refers to articles 12 to 15, does he intend for a future article 13a to be included in that reference? One could argue that this is always an issue when a law is changed, as any part of the law is a set similar to a range. Chapter 2 is a range including several articles, and when a new article is inserted into Chapter 2, this means that any references to Chapter 2 now also include this new article. Still, named sections like Chapter 2 have a clear theme, and it seems less likely that a new article inserted into a chapter should not be included in a (earlier) reference to that chapter.

Another group of references that do not include a specific index are those that refer to the *previous* article or subparagraph, etc. Like ranges, these require a derivation to determine the target. Also like ranges, modifications may unintentionally change the target of the reference. For example, article 61 of the Military Penal Code refers to *the previous article*, meaning article 60. In 2000, a new article 60a was inserted between articles 60 and 61, without modifying article 61[59]. As a result, the (literal) target of the reference in article 61 has changed to article 60a. In both situations (ranges and the use of *previous*), it may be that the resulting situation is

[59] Act of June 2nd, 1999 (Stb. 1999, 343), enacted on January 1st, 2000.

in fact desired and correct, but it seems that using explicit references (i.e. those that contain the index of each target) is less prone to produce errors.

A possible next step for the detection of references is to detect the type of reference as well. As mentioned in the introduction, references may serve various goals, such as referencing a target in order some definition or to modify it. Adding such a type to each reference will make it more valuable for navigating these links between documents.

Next to improving access to legislation, an application that can automatically detect references in legal sources opens up the possibility to perform network analysis of the corpus of sources of law. Such research has already been performed on the French legal codes by Mazzega, Bourcier and Boulet (2009), on court decisions of the United States Supreme Court, by Smith (2005) and Bommarito, Katz and Zelner (2009), on the Estonian Law of Obligations by Liiv, Vedeshin, and Täks (2007) and Dutch Supreme Court cases by Winkels and de Ruyter (2011). Such research can help identifying important sections of the law and key cases, as well as detect clusters of related sources. This, in turn, may lead to improved search techniques for legal sources and new ways to measure complexity of laws (see Bourcier and Mazzega, 2007).