Searching References to Secondary EU Legislation

Marc van Opijnen[1]

[1] Council for the Judiciary, the Netherlands
mvanopijnen@rechtspraak.nl[1]

Abstract. At the national level numerous judgments are referring to secondary EU legislative instruments, like directives and regulations. When searching for these citations however the user is confronted with low recall and precision, due to inconsistent citation practices and poorly configured search engines. After an in-depth analysis of these causes this paper describes a language- and platform neutral solution. A critical note is added on EU legislative custom and the EUR-Lex standard for referencing elements of secondary EU legislation.

Keywords: Structuring legal documents, EU legislation, case law search.

1 Introduction

The national judge plays an important role in upholding the European legal order, which becomes increasingly interwoven with the national legal systems of its member states. Not only does he have to apply and interpret the EU treaties but also the secondary EU legislation: tens of thousands of regulations, directives, decisions and recommendations.

Apart from having access to these EU acts themselves and the jurisprudence of the Court of Justice – as supplied by the EUR-Lex website – judges and other legal professionals need access to databases with national court rulings on Community law. Improving cross-border access to national case law is listed as a priority in the European e-Justice action plan[2] and stressed by the European Parliament in its resolution *on the role of the national judge in the European judicial system.*[3]

As will be demonstrated in § 2 though users encounter severe problems searching databases for judgments on EU law: case law is generally stored as plain text, without well-structured references or metadata. Due to poor adherence to citation standards by authors, search engines misunderstanding punctuation marks in queries, and non-reprehensible ignorance by users, recall and precision on searches for documents citing secondary EU legislation are dissatisfyingly meagre. In multilingual and/or distributed environments these problems are multiplied.

---

[1] Thanking Freek Leemhuis and Anita van Dijk for software engineering.

[2] OJ 2009/C 75/01

[3] http://www.europarl.europa.eu/sides/getDoc.do?type=TA&language=EN&reference=P6-TA-2008-0352

The solution described in § 3 – pattern matching in combination with converting document numbers to Celex-numbers – leads to a radical improvement in search results. How to search for specific elements of EU acts is described in § 4, together with a proposal for the revival of a European standard. Conclusions are drawn in § 5, together with some recommendations for future work at the European level.

This paper is based on a Dutch research project. For reasons of understandability and to demonstrate the flexibility of the solution, examples are also drawn from British and German databases; code examples are translated into English. Dutch judgments indicated with '*' can be found by LJN on http://zoeken.rechtspraak.nl, UK judgments indicated with '**' by neutral citation number on http://www.bailii.org.

It should be stressed that the scope of this project is to facilitate searching case law by cited EU acts; the design of a referencing standard is not intended. It should also be kept in mind that the scope is limited to secondary EU legislation. The primary treaties are cited in a completely different way, have a different Celex-numbering system and face additional problems with renumbered articles. They need a different approach which cannot be discussed in this paper.

2 Searching for References to Secondary EU Legislation

Although legal citation guides at the national [1] or European level [2] prescribe how secondary EU Legislation should be cited, daily practice shows a range of variants. As an example we search for case law on 'Regulation (EEC) No. 1408/71 of the Council of 14 June 1971 on the application of social security schemes to employed persons and their families moving within the Community'.

Although 534[4] judgments in the public database of the Dutch judiciary cite this regulation, only five of them use the full title. Instead of using the full title, [1] advises the (Dutch) abbreviation: *"Verordening (EEG) nr. 1408/71 (PbEG 1971, L 149/71)"*, the last part being a reference to the Official Journal of the EU.

Table 1. Examples of citation styles from the Netherlands, UK and Germany.

Netherlands	United Kingdom	Germany
Verordening (EEG) 1408/71	Regulation 1408/71	Verordnung Nr. 1408/71
Verordening (EG) nr. 1408/71	Regulation (EEC) No 1408/71	Verordnung (EWG) Nr. 1408/71
EG-Verordening nr. 1408/71	Regulation (EC) 1408/71	Verordnung (EWG) 1408/71
EEG-Verordening nr. 1408.71	Regulation 1408/71 EEC	Verordnung [EWG] 1408/71
EG-Vo. 1408/71	Regulation No 1408/71	VO (EWG) Nr. 1408/71
Vo 1408/71	EEC Regulation 1408/71	VO (EWG) 1408/71
verordening nr. 1408/71	Regulation EEC No. 1408/71	VO <EWG> 1408/71
EG VO 1408/71	EEC Reg. 1408/71	EWGV 1408/71
Vo. 1408/71 (EG)	Reg. EEC 1408/71	EWG-VO Nr. 1408/71
Verordening 1408/71/EEG	Reg. 1408/71	EWG-VO 1408/71

[4] Counted on 7-7-2010, using the methodology described in this paper.

There is not one single judgment though using this prescribed abbreviation, but 194 judgments (36.3%) use this notation at least once without the reference to the Official Journal. The vast majority of judgments use a whole range of other notations. The same can be witnessed in other languages. In table 1 some examples from the Netherlands, UK and Germany are collected.

From these examples and an analysis of search engine behaviour on various websites we can learn:

1. There is a wide range of spelling variants. Therefore the number of queries to be written to obtain an acceptable recall is nearly unlimited.
2. Notwithstanding the existence of clear rules on the use of 'EEC' (European Economic Community) and 'EC' (European Community) for legislation entered into force before, respectively after 1 November 1993 (date of entry into force of the Maastricht Treaty), citing authors seem to be ignorant of the difference. Even in one judgment both abbreviations are used to indicate the same regulation (e.g. LJN AE3898*).
3. Using only the document number of the regulation for search ('1408/71') is risky because the fourth Dutch example shows that other punctuation marks than slashes might be used - dots and hyphens are quite regular. Moreover, the last Dutch example shows that the notation of the document number that is prescribed for i.a. directives is sometimes erroneously also used to cite regulations. Therefore also this last example might not be found when searching on '1408/71'.
4. Slashes (or other punctuation marks) in search queries might lead to unexpected and unwanted results, because they are often used for specific search instructions, or just not understood by a search engine. This can be illustrated[5] by searching Bailii for '1408/71' on all but the European case law databases. 235 Results are returned, of which only 128 indeed cite Regulation (EEC) No 1408/71, leaving precision at 54%. The other documents just have '1408' in title or text, referring to anything but the EEC-regulation.
5. Using (only) document numbers also increases the risk of poor precision because document numbers are ambiguous at the European level itself. A search for cases citing 'Directive 2001/7/EC' using just '2001/7/EC' might result in documents citing 'Commission Decision 2001/7/EC' instead.
6. The use of wildcards to find erroneous citations is too difficult for the average user – if possible at all. How to formulate a wildcard-query to find both '2001/7/EEC' and '2001/7(EC)' without retrieving '2001/78'?
7. According to [1] and [2] the year-part in the document number has to be written in two digits if the act is concluded before 2000, but in four digits if it is concluded in or after 2000. Non-compliance is easy to find: LJN AR3078* cites "Verordening (EG) nr. 2419/01" and LJN AT7248* cites "Verordening (EG) nr. 2316/1999". To improve recall queries should thus be formulated in two-digit and four-digit variants.
8. Document numbers for regulations are written as [number][year], while the document numbers of directives, decisions and recommendations are constructed in reverse order: [year][number]. In document numbers where the number could

---

also be a year, searching without knowledge on the type of legislative document impairs precision. E.g., searching on '96/95' might lead to documents citing Commission Regulation (EC) No 96/95 (no 96 from the year 1995), Council Directive 96/95/EC (no 95 from the year 1996) or Commission Decision 96/95/EC (also no 95 from the year 1996).

In multilingual and/or distributed environments these problems are multiplied, which can be illustrated by taking the perspective of the user of the 'Common Portal of Case Law'.[6] This metasearch engine, querying nineteen databases from eighteen EU countries was developed by the Network of the Presidents of the European Supreme Courts , to meet the need for comparative studies on the national implementation of EU law. To facilitate keyword searches the multilingual thesaurus Eurovoc was integrated in the user interface, but most questions on the interpretation of EU law are centered around a specific (element of a) legal instrument, and not on hard-to-define legal concepts. Most searches thus are performed on the document numbers of EU legal acts. With the abovecited problems of national databases it is easy to understand the poor results of the Common Portal: every country has its own abbreviations and notation habits, and every database its own particularities in handling punctuation marks.

3 Solution

So, if a legal researcher wants to know which judgments are rendered on Regulation (EEC) No 1408/71, how can he be served with the best possible recall and precision?

Query-expansion is not a good idea: it requires a very complex script, a lot of resources during runtime, and it won't solve all problems mentioned. Instead, a solution was developed that doesn't try to improve the search engine, but to help it by structuring the indexed documents. It has three basic ingredients: pre-processing with pattern matching (§ 3.1), conversion of document numbers to Celex-numbers (§ 3.2) and structured storage (§ 3.3).

3.1 Pattern Matching

When citations are stored in a structured way, searches can be performed much faster. To achieve this they have to be recognized first. Regular expressions are the most flexible way to do this.

For Dutch document texts the regular expression for recognizing the citation of an EC-regulation ('Verordening') is:

```
(V|v)(er)?(o|O)(rd)?(ening(en)?|\.)?\s*((?i)(n(umme)?(r
|o)\.?)\s*)?\d{1,4}(/|\.)(19|20)?\d{2}(/|-
|\s)\(?(E(E)?G(\sen\s|,\s|/)Euratom|E(E)?G|Euratom)\)?

(E((E)?G|uropese)(-
```

```
|\s+))?(V|v)(er)?(o|O)(rd)?(ening(en)?|\.)?\s*((?i)(n(u
mme)?(r|o)\.?)\s*)?\d{1,4}(/|\.)(19|20)?\d{2}
|
(V|v)(er)?(o|O)(rd)?(ening(en)?|\.)?(\s*|-
)\(?(E(E)?G(\sen\s|,\s|/)Euratom|E(E)?G|Euratom)\)?,?(\
s*|/|-
)?((?i)(n(umme)?(r|o)\.?)\s*)?\(?\d{1,4}(/|\.)(19|20)?\
d{2}\)?
```

The version for English texts is:

```
(R|r)eg(ulation|\.)?\s+((?i)(n((umbe)?r|o)\.?)\s*)?\d{1
,4}(/|\.)(19|20)?\d{2}(/|-
|\s+)\(?(E(E)?C(\sand\s|,\s|/)Euratom|E(E)?C|Euratom)\)
?
|
(E((E)?C|uropean)(-
|\s+))?(R|r)eg(ulation|\.)?\s+((?i)(n((umbe)?r|o)\.?)\s
*)?\d{1,4}(/|\.)(19|20)?\d{2}
|
(R|r)eg(ulation|\.)?(\s+|-
)\(?(E(E)?C(\sand\s|,\s|/)Euratom|E(E)?C|Euratom)\)?(\s
+|/|-
)?((?i)(n((umbe)?r|o)\.?)\s*)?\(?\d{1,4}(/|\.)(19|20)?\
d{2}\)?
```

Important features of this regular expression are:

1. It catches all regular spelling variants for 'Regulation' / 'Verordening', including all abbreviations.
2. As is shown in the examples the E(E)C-part is often left out, and therefore it is not mandatory in the regular expression. The word 'Regulation' / 'Verordening' or an abbreviation thereof is required though.
3. Regulations of EC, EEC, Euratom, and combined E(E)C – Euratom regulations are caught, regardless of whether E(E)C is spelled correctly.
4. It catches both two-digit and four-digit variants of the year.
5. It catches all (thinkable) orders of the constituing parts of the citation, i.a. all those listed in the examples.

Comparable regular expressions were written for decisions, recommendations and directives. They are not displayed here because of their similarity, although they have the number and the year in reverse order. A regular expression was also written for framework decisions (and joint actions), which are so specific in their document numbering that the regular expression is quite simple:

```
(19|20)?\d{2}(/|\.)\d{1,4}(/|-|\s)?(JBZ|JHA)
```

Although these regular expressions are developed on vast numbers of examples, the inventiveness of legal professionals to come up with new variants is nearly infinite – small adaptations might therefore still be necessary. Adaptations due to the Lisbon Treaty are also not implemented yet.

3.2 Normalizing to Celex-numbers

The Celex-numbering system is the unique numbering system for European legal documents. Because of its strict syntax it's more suited for electronic environments than the document number. Moreover, the Celex-numbering system is basically language-neutral. It is used by the EUR-Lex website where the URI-system allows for addressing any specific language version.

Although there are some refinements, for our goal the basic architecture of the Celex-number suffices. It consists of four basic parts, which we apply here for Regulation (EEC) No 1408/71:

- Sector, which is for all secondary EU legislation: '3';
- The year, always written in four digits: '1971';
- Type of document, which is 'R' for Regulation;
- Ordinal number, which is the ordinal part of the document number, if necessary with leading zeros to form a four-digit number: '1408'.

So, the Celex-number for our regulation is '31971R1408'. Hence, the conversion from the document numbers we found with the regular expressions to Celex-numbers is rather straightforward. It is depicted in the UML activity scheme of figure 1.
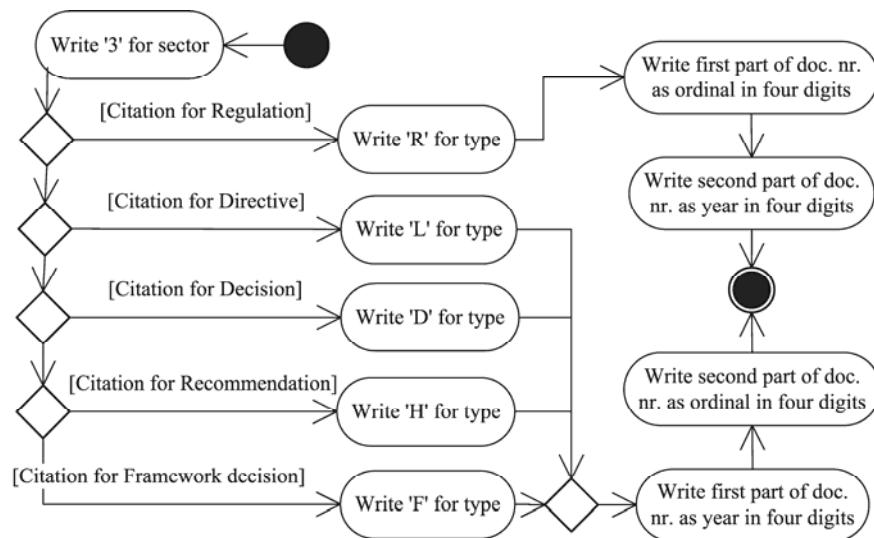
Fig. 1. Decision scheme for constructing Celex-numbers from outcome of regular expressions.

Finally, the Celex-number is validated against the EUR-Lex repository. If the constructed Celex-number turns out to be non-existent, it can be ignored or put in a queue for manual processing.

3.3 Storage

To facilitate search – but also e.g. deeplinking – the Celex-number has to be stored in a structured way. This can be done in the text itself, or in metadata fields. With XML markup a text[7] would read:

```
<text>(...) within the field of social security covered
by <citation Celex="31971R1408">Regulation No.
1408/71</citation> the legislation at issue
(...)</text>
```

Of course the Celex-number could also be stored as RDF or translated to referencing standards like URN:LEX[8] or CEN/Metalex.[9]

3.4 Additional Challenges

Although the method described works quite well, some challenges remain.

Firstly, regulations and directives are sometimes cited by a global or local alias. A local alias is declared in the document itself and is only used locally. An example might be: *"Regulation (EEC) No. 1408/71 (hereafter referred to as 'the Regulation') (...) Article 14 of the Regulation describes the exceptions."*

By having the parser look for phrases like 'hereafter referred to' in conjunction with words like 'Regulation', 'the Regulation' in the second sentence is recognized as Regulation (EEC) No 1408/71. A search on article 14 of this regulation will therefore be successful – after also § 4 (on searching for specific elements) is implemented.

Global aliases pose other problems. Searching is seriously hampered if a global alias is not also declared as a local variable. This can be illustrated with [2009] EWHC 906 (Ch)**, where the 'Seventh Company Law Directive on Consolidated Accounts' is cited. If one searches in Bailii on (variants of) 'Directive 83/349/EEC', this judgment will not appear in the result list, because the judge only used the global alias to cite it. The only solution is to cater for a list of known aliases to scan the text in the parsing process and to add mark-up like they were document numbers. The list requires manual maintenance because global aliases are mostly not official. Sometimes even more aliases exist for one act – e.g. 'Services directive' and 'Bolkestein directive' for Directive 2006/123/EC.

A second challenge is in the fact that EU instruments adapting or implementing other EU instruments also contain the document number of the adapted or implemented instrument. If a citation reads: '(...) Commission Regulation (EC) No 120/2009 of 9 February 2009 amending Council Regulation (EEC) No 574/72 laying down the procedure for implementing Regulation (EEC) No 1408/71 on the application of social security schemes to employed persons, to self-employed persons and to members of their families moving within the Community (...).' our parser will translate it into:

---

[7] All examples in this and the next paragraph are taken from [1998] UKSSCSC CIS_863_1994**

[8] http://tools.ietf.org/html/draft-spinosa-urn-lex-00

[9] http://www.metalex.nl/

```
<text>(...) <citation Celex="32009R0120">Regulation
(EC) No 120/2009</citation> of 9 February 2009 amending
Council <citation Celex="31972R0574">Regulation (EEC)
No 574/72</citation> laying down the procedure for
implementing <citation Celex="31971R1408">Regulation
(EEC) No 1408/71</citation> on the application of
social security schemes to employed persons, to self-
employed persons and to members of their families
moving within the Community (...)." </text>
```

In other words: three regulations are recognized, while only one regulation is actually cited. One solution could be to scan the judgment for official titles. Given the number of more than 100.000 existing acts this requires a lot of hardware resources. On the other hand the problem is a minor one: as was established in the first paragraph just about 1% of citations are made using official titles. Moreover, legal writers usually cite consolidated versions only, limiting the problem to the implementing acts.

4 Citing Specific Elements of Secondary EU Legislation

Judgments not only refer to EU acts as such, but also to specific elements, e.g. 'article 4'. A script was developed to recognize these elements, but because it is quite lengthy and specific for the Dutch language and citation style, it will not be discussed here – comparable techniques have been described elsewhere [3]. The result is stored with specific attributes, like in:

```
<text>(...) In Hoeckx a benefit like the minimex was
held not to be covered by <citation Celex="31971R1408"
partType="article" partNumber="4">Article 4 of
Regulation 1408/71. (...)</citation>. </text>
```

In a national environment this solution works well, but when searches have to be performed on a variety of (international) databases, problems will arise because of different naming and design rules. This is even more evident in more complex citations like: *"(...) as an aid in interpreting the ambiguous English text I have also considered the French text of Article 4(2a)(a) of Regulation 1408/71."* This citation could be referenced in many ways. To give just some examples:

```
<citation Celex="31971R1408" partType="article"
partNumber="4" partType="paragraph" partNumber="2a"
partType="listitem" partNumber="a"/>

<citation Celex="31971R1408" article="4" paragraph="2a"
listitem="a"/>

<citation Celex="31971R1408" part="article 4(2a)(a)"/>
```

It goes without saying that a distributed or federated search will be hampered seriously by such a variety of syntaxes. Therefore, as long as documents are not available in a standardized format or with a standardized referencing mechanism, it would be advisable to use – at least – the referencing method of EUR-Lex itself.

Advantages of this system are its language-neutrality, the short and straightforward notation, and the fact that it is already used in the EUR-Lex system. Unfortunately, it also has some disadvantages. Firstly, it is not documented on the EUR-Lex website, and only very poorly described in [4]. Secondly, it doesn't seem to be used consistently anymore in EUR-Lex, possibly because it is not a very user-friendly notation; the advantages of having a machine readable standard are not imminent for the EUR-Lex site itself, since deeplinking specific elements of EU acts is not possible.

To demonstrate its use, the standard is 'reverse engineered' here – though probably not completely. Examples can be witnessed on the metadata page of Regulation (EEC) 1408/71,[10] especially under the heading 'Affected by case'. The main constituting elements are listed in table 2. Some examples are listed in table 3.

Table 2. Basic elements types used in EUR-Lex for referencing parts of secondary EU legislation.

Prefix	Meaning
A, with the number in at least two digits	Article
P	Paragraph
N	Annex
L	Listitem or sub-listitem
	Point
TIT	Title
CH	Chapter

Table 3. Examples for defining elements of secondary EU legislation using the EUR-Lex referencing system.

Example	Meaning
A06	Article 6
N2	Annex 2
A10P1	Article 10, paragraph 1
A04P1LB	Article 4, paragraph 1, listitem B
A01LALII	Article 1, listitem A, sub-listitem II
N5IP1	Annex 5, point I, paragraph 1
A71P1LBLII	Article 71, paragraph 1, listitem B, sub-listitem II
N5CP1LB	Annex 5, point C, paragraph 1, listitem B
TIT3CH1	Title 3, chapter 1

In references the element is suffixed to the Celex-number.[11] As a result, one might be inclined to conclude that our example-reference to article 4(2a)(a) of Regulation (EEC) No. 1408/71 would read:

```
<citation Celex="31971R1408-A04P2ALA"/>
```

But unfortunately this is not correct. This is caused by the fact that while the EUR-Lex referencing system is language-neutral, the numbering of elements is not. Comparing the various language versions of § 3.3 of [2] reveals the official EU-policy to use country specific formats for numbering elements in a legal act – differences become especially apparent in elements inserted by amendments. Where the abovecited text refers to 'Article 4(2a)(a)' in (i.a.) the English, Polish, Finnish and German versions of the regulation, in (i.a.) the Dutch, Spanish, Italian and French version the paragraph-number is not '2a' but '2 bis'. Because the EUR-Lex referencing system uses the French style[12] the correct reference is "31971R1408-A04P2BISLA" instead of "31971R1408-A04P2ALA".

It goes without saying that cross-border searches are complicated by this legislative anomaly: when the references are constructed, a conversion table between the language of the document and French would have to be used.

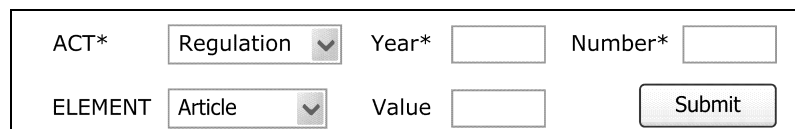## 5 Conclusions and Future Work

The solution described in this paper is implemented in a testing environment on a case law database with 770.000 documents.

To test the recall the database was searched for all documents containing the strings "1408" and also the string "71". 1204 documents were returned. These documents were parsed using the parser described in § 3. In 1076 documents a reference to Regulation (EEC) No. 1408/71 was recognized. Manually testing the remaining 128 documents revealed that 120 documents coincidentally contained both strings without referring to the European regulation, two documents contained typo's in the reference to the regulation and six documents were discarded erroneously. Recall can therefore be calculated at 99,3%.

Analyzing the text strings which were recognized by the parser showed that no strings were recognized that did not refer to Regulation (EEC) No. 1408/71, leaving precision at 100%.

Apart from local aliases the 1076 documents contained 3382 references to the regulation, of which only 450 (13,3%) were of spelling prescribed by [1]. In these 3382 references 49 different spelling variants were used.

For the end-user a simple search interface was developed as depicted in figure 2 – to query the database for judgments citing specific EU acts

| ACT* | Regulation ▾ | Year* | | Number* | |
| ELEMENT | Article ▾ | Value | | Submit | |

Fig. 2. User interface to search for case law on EU secondary legislation.

---

[12] Although this is not publicly documented.

Type, year and number of the act are mandatory, while element type and value (e.g. 'article 4') are optional. From the first three fields a Celex-number is calculated, which is used to build the final query. In a production environment the search could be combined with criteria like type of court and date of judgment. Although the interface is not implemented in a production environment yet, user feed-back has been very positive: research is foreseen to be reduced from hours (or even days) to seconds.

With small adaptations in the regular expression for recognizing the legal acts this solution can be implemented in any software environment, in any EU member state. To recognize the cited elements additional scripts have to be written, focussed on country-specific citation practice.

Limitations have to be kept in mind though: the system is limited to secondary EU legislation, it is not meant to be a referencing system and does not feature searches for elements as valid on a specific date. Nevertheless, given the fact that most national case law databases lack any facility to search for cases implementing or applying secondary EU legislation, this solution can offer immediate and substantial improvements for end-users. Moreover, when applied in various member states, it could facilitate cross-border search tools of which the Common Portal of Case Law was already mentioned; others are discussed in [5].

Meanwhile this study revealed some weaknesses in EU legislation and publication practices.

Firstly, using language-specific numbering formats in secondary EU legislation, as prescribed by [2], seriously hinders EU wide legal research and needlessly complicates interoperability initiatives. Moreover, this legislative practice could be in defiance of resolutions of the European Parliament. The aforementioned resolution on 'The role of the national judge in the European judicial system' *"Insists on the need for (…) greater terminological coherence between legal instruments."[13]* It is well arguable that this also covers syntactical coherence. Moreover, in the E-Justice resolution[14] the European Parliament calls upon the Commission to *"set up suitable machinery to ensure that all future legislation (…) is designed in such a way that it can be used in on-line applications."*

Secondly, the EUR-Lex standard for referencing specific elements of secondary legislation should be revitalized and publicly documented. Albeit only for better comprehension of the hundreds of thousands of these references on the EUR-Lex website on the interpretation of which the users are left in the dark.

References

1. Bastiaans, M.H. (ed.): Leidraad voor juridische auteurs 2004. Kluwer Deventer (2004).
2. Interinstitutional style guide, Vade-mecum for editors. Publications Office for the EU, Brussels (1998).
   Updated on-line version at http://publications.europa.eu/code/en/en-000100.htm.

---

[13] See footnote 3, § 33.
[14] http://www.europarl.europa.eu/sides/getDoc.do?type=TA&language=EN&reference=P6-TA-2008-0637

3. Maat, E. de, Winkels, R, Engers, T. van: Automated Detection of Reference Structures in Law. In: Engers, T. van (ed.): Legal Knowledge and Information Systems – Jurix 2006: The Nineteenth Annual Conference. IOS Press, Amsterdam (2006).
4. Celex's Technical Manual for Hosts. European Commission, Brussels (1991).
5. Opijnen, M. van: Finding Case Law on a European Scale - Current Practice and Future Work. In: Francesconi, E., Sartor, G., Tiscornia, D. (eds.): Legal Knowledge and Information Systems – Jurix 2008: The Twenty-First Annual Conference. IOS Press, Amsterdam (2008).