

*Chapter Nine*

.....

## Some hints on searching CELEX

To a human being, a word is a variable quantity. Words have a relationship to each other and are organized in order to convey a specific meaning. To a computer, on the other hand, a word is simply a sequence of characters, which begins and ends with a known specified character, such as a blank space. To a computer, a word is always the same. Thus, there is a contrast between the flexibility of the human mind and the rigidity of the computer.

When text is loaded into a computer, it has to conform to the constraints of the machine. When the human user wishes to search that text and abstract the relevant parts of it, the output of the machine has to conform to human requirements. These processes of conversion and interpretation between the machine and its human counterparts are essential ingredients in a computerized retrieval system.

A useful starting point is to consider the processes involved in storing text in a computer for subsequent retrieval. The first thing a computer system does when given text is to arrange the words of the text into some logically ordered sequence, alphabetic order for example, which can be used subsequently for searching purposes. In addition, since there is a relationship between the individual words in the text, it is necessary to record and store the exact position of every occurrence of each word in the text. If this is done for every individual word used in the text, then these computer files are several times larger than the original text. Fortunately, there are ways of reducing this to a manageable size.

A feature of printed text is that it is always organized in component parts, that is to say, the text has structure. This structure may simply be that of a book, which has chapters, paragraphs, and sentences. However, the texts in legal documents are structurally more complex. A case report will normally have a title, a headnote, a list of citations, the judgment and other sections, which contain the names of the judge(s), the counsel and the solicitors in the case. This structure is very useful for quicker and more efficient retrieval, since a search can be restricted to particular parts of a document.

Legislation has a completely different structure from case reports since it consists of Acts, which have sections, sub-sections, schedules, etc. CELEX contains not only legislation and case law, but also treaties, proposals, national implementation references and parliamentary questions. Even though each type of document has its own structure, there is a good deal of common ground between them. This structure contains valuable information that is useful to the end user and can significantly help the search process.

CELEX is a hybrid database, that is, it consists of documents that are a mixture of free text and highly structured, codified information. The simplest approach to searching CELEX is to treat it as a full-text database and to use words to locate documents.

CELEX contains some unusual terminology because much of its text has been translated from other European languages, mainly French, so it is often sensible to begin a search by using only a single word. If necessary, the single word search may then be refined by using additional words and combining them with operators such as **AND**, **OR** and **NOT**. Here are some examples:

**Example 1**

- *air AND pollution AND directive*

This will find those CELEX documents containing all of these words.

**Example 2**

- *air OR atmosphere OR marine*

This will find those CELEX documents that contain any one of these terms.

**Example 3**

- *air NOT atmosphere*

This will find those CELEX documents that contain the word "air" but not the word "atmosphere".

The **NOT** operator is useful for excluding irrelevant documents. It is very powerful and should be used with great care. Brackets should always be used to make the application of **NOT** absolutely clear. For example:

- *(air NOT atmosphere) AND pollution*
- *(man OR woman) NOT (man AND woman)*

The simplest use of **NOT** is with single words:

- *Commission NOT Council*
- *Waste NOT dangerous*
- *Regulation NOT directive*

Note that search terms may occur anywhere within a document. If the document is a lengthy one, the search will still be positive if the term "air" occurs in the first line of the text and the term "pollution" in the last line, but in this case the document is unlikely to be particularly relevant. CELEX is a very large database containing thousands of long documents, so it is often necessary to revise searches to make them more specific. One way is to add more search terms, for example:

- *air AND pollution AND directive AND motor AND vehicle*

This use of **AND** to combine a series of terms is probably the most common and effective form of searching. It is good practice to put the lowest frequency word first, since this will lead to a faster response in most retrieval systems.

A further refinement is to specify the relative position of the terms. Most retrieval systems include this sort of facility. In Justis.com, examples of proximity searching are:

- *air WITHIN 25 OF pollution* (air must be within 25 characters of pollution)
- *air WITHIN 40 AFTER pollution* (air must be within 40 characters of pollution, and must be after pollution)
- *air WITHIN 10 BEFORE pollution* (air must be within 10 characters of pollution, and must be before pollution)
- *air NEAR pollution* (air must be within 40 characters of pollution. It is the same as air within 40 of pollution)

The use of a phrase is even more precise:

- *air pollution*
- *motor vehicle pollution*

CELEX can be interrogated in much the same way as any other full-text database, provided that the problems of “Eurojargon” and translation are kept in mind.

The first step when using CELEX is to decide which Sector is most likely to contain the answer to the question being asked, for example:

- *Does the question concern a Treaty?* → Choose **Sector 1/2**
- *Does the question concern Legislation?* → Choose **Sector 3/4**
- *Does the question concern future Legislation?* → Choose **Sector 5**
- *Does the question concern a case?* → Choose **Sector 6**

Of course, the question could involve more than one of these document types, so the complete answer may require searching within all the relevant sectors. This is particularly true of the legislation held in **Sector 3/4**, since there may be proposals for new legislation in **Sector 5**, which will have to be found in order to present a complete picture. However, it is better to search sector by sector, in order to be clear in developing the search. Once the sector has been selected, the question should be examined to see if use could be made of the field structure that has been superimposed on the text.

Consider, for example, the following question:

- *“Find the directives which are concerned with the regulation of banks, building societies, and other financial institutions. Are there any proposals for new regulations concerning them?”*

The question may be developed as follows:

### **Step 1**

Select **Sector 3/4** for the legislation.

### **Step 2**

Enter the following search:

- *bank\* AND directive*

The answer returned is: “+900 documents satisfy this query”. This number is too large to manage but it is still useful to inspect a few titles and look at the vocabulary which is used. Notice that the phrase “credit institutions” occurs frequently; this is Eurojargon for banks, building societies, etc.

### **Step 3**

Rephrase the question as:

- *credit institutions AND directive*

The answer returned is: “+150 documents satisfy this query”. This is an easier number to handle, but is still large. A look at the titles will show that a number of regulations and decisions have been found, in addition to directives.

### **Step 4**

Rephrase the question by making use of the field structure:

- *credit institutions AND FORM[directive]*

The answer is: "+50 documents satisfy this query". All of these documents are directives. With practice, this question will become the first step taken.

The second part of the original question asks if there are any proposals on this subject, so proceed as follows:

### **Step 1**

Select **Sector 5** for proposals.

### **Step 2**

Make use of the field structure again and submit the following query:

- *FORM[proposal] AND directive AND credit institutions*

The answer is: "+60 documents satisfy this query".

Now there is a complete answer to the original question. It pays to become familiar not only with the structure and the layout of the documents, but also with the style of drafting and the terms used. Drafting is often done in committees of experts and not by legal draftsmen. The translation process also adds another layer of difficulty, and this must be kept in mind when searching, since nouns used as adjectives are not permitted in French: phrases such as "product liability" or "banking directive" are not used. It is safer to use the **AND** operator, get some answers, and then refine the search question with proximity operators. Then no relevant document will be missed. For example:

- *product AND liability*
- *product NEAR liability*

Furthermore, it pays to become familiar with the terminology used. As mentioned above, "credit institutions" covers banks, building societies, and other financial organisations. Another example is the word "concentration", which means the merger or takeover of commercial undertakings.

Some of this jargon is covered in the **EUROVOC** thesaurus (<http://europa.eu.int/celex/eurovoc>), which is by no means comprehensive, but it is helpful. The EU is developing extremely fast and new words are continuously being introduced. It really is necessary to read actual examples of CELEX documents, however boring they may be.

It is also worth spending a little time studying the **DOCNUM** field to see how the unique document number is created (see **Chapter 3**). Document numbers are used in all the cross-referencing fields, which are used to link related documents. The fields **LEGBASE** (legislative base), **LEGCIT** (legislation cited), **MODIFIES** (legislation amended) and **MODIFIED** (legislation amending) all use the document number. These fields must always be checked in any retrieved document to see whether the current document has any effect on earlier documents or whether it has been amended by a later one. A single search using the **MODIFIES** field will find all the amendments to a particular document, specified by its document number.

The **FORM** field is very useful for limiting a search to one particular kind of document. For example, *FORM[directive]* will limit the search to directives only. This is an efficient way of searching CELEX since there are nearly 100,000 pieces of legislation stored in **Sector 3/4**.

The **DATE** fields are useful in limiting searches to a particular period of time. The most reliable is the **DOC** field, which uses the ISO form *yyyy/mm/dd* for the date. For example, *DOC[1990/05]* will find all documents signed in May 1990.

The **SUB** field uses a number of broad descriptions, each covering an area of Community law. Use of this field will limit the search to a given subject area, which may then be searched efficiently in greater detail. A full list of these descriptions is given in **Appendix 4**. Examples are:

- *SUB[free movement of capital]*
- *SUB[social provisions]*
- *SUB[internal market]*

Lists of CELEX fields are given in **Appendix 1**. Familiarity with these fields will enable better use to be made of the CELEX database.